MARBLE: Music Audio Representation Benchmark for Universal Evaluation

Abstract

Music understanding covers many tasks of extracting information from audio and analyzing it, but the small size of available datasets constrained by copyright and annotation costs limits model performance, and music model evaluation yet lacks a unified benchmark, and existing evaluations are fragmented, making it difficult to comprehensively compare different techniques. Therefore, this replication study focuses on the paper "MARBLE: Music Audio Representation Benchmark for Universal Evaluation" and aims to replicate and analyze the MARBLE project, a comprehensive music audio Representation Benchmark for evaluation of various Music Information Retrieval (MIR) tasks. This replication mainly focuses on ten tasks on six publicly available datasets that have been open-sourced in the paper's source code, to complete all the data processing in the pre-processing stage and to provide a comprehensive evaluation of the two pre-trained models. This paper will present a detailed comparative analysis of existing techniques, methodology, and experimental results of the reproduction process.

Keywords: Benchmark, Evaluation, Music Information Retrieval, Music representation.

1 Introduction

In this day and age, the cross-fertilization of AI with the arts is becoming increasingly widespread, with applications such as image generation and novel co-creation flourishing. However, the application of AI in the field of music, especially music understanding, is still at a relatively early stage. Music understanding covers a series of tasks to automatically extract information from raw music audio, such as music classification, emotion recognition, pitch estimation, and analysis of musical features such as rhythm, melody, and harmony. However, the size of labeled music datasets is usually small due to factors such as copyright and labeling cost, which largely constrains the performance of supervised models in music understanding tasks.

Self-supervised learning (SSL) has shown great potential in a variety of tasks (e.g., in the fields of natural language processing and computer vision) with limited labeled data, which has led to the emergence of research related to SSL-based learning of audio representations and music pretraining models. Nevertheless, current evaluations of music models lack comprehensive and unified benchmarks, and existing evaluations are fragmented, making it difficult to provide objective comparisons and insights across different technologies. For example, in existing studies, SSL music systems are mainly evaluated by downstream task datasets covering tasks such as genre classification, emotion classification, instrument classification, music labeling, key

detection, music detection, beat tracking, and cover song detection; however, the studies usually adopt different experimental setups, and there are fewer explorations of sequential tasks such as beat tracking and source separation.

To address these issues, we set out to replicate the MARBLE benchmark proposed in the paper.MARBLE [1] aims to fill the gaps in the existing evaluation system of music models by defining a comprehensive four-layer taxonomy (comprising acoustic, performance, score, and high-level descriptions) for organizing and evaluating music information retrieval tasks. The original paper establishes a unified evaluation protocol based on 12 publicly available datasets covering 18 downstream tasks, providing a standardized, generic and easy-to-use framework for music model evaluation.

2 Related works

With the continuous development of artificial intelligence technology, music understanding and assessment research results in natural language processing, computer vision and other fields have gradually emerged. As a key component of Music Information Retrieval (MIR) and AI applications, Music Understanding Models and Music Evaluation Benchmark, with the help of computer vision, Natural Language Processing (NLP), and machine learning technologies, researchers are able to realize multi-dimensional understanding of music, further promoting the deep integration of music and technology.

2.1 Music Understanding Model

The music understanding model is designed to allow computers to 'understand' music and perform a variety of analysis and reasoning tasks, such as pitch recognition, rhythm analysis, emotion inference, melody extraction, etc. The model requires not only the analysis of the audio signal, but also the analysis of the audio signal. It not only needs to parse the audio signal, but also needs to consider multimodal information such as lyrics, video, emotion, etc. In recent years, with the breakthrough of deep learning technology, Long Short-Term Memory (LSTM) [2], Transformer [3] and other technologies have been used in a large number of applications, especially the addition of the large language model, which has made significant progress in the music understanding model.

A very important step in understanding the music is the effective representation of the music signal. Traditional methods, including analysis based on low-level features such as notes and clefs, are not sufficient to support model training. The advent of deep learning has made it possible to learn high-level features of music by learning them. For example, audio signals are converted to 2D images using time-frequency maps and fed into CNNs to extract information such as rhythm and melody [4]. In addition, RNN and LSTM are widely used to capture temporal information in music, which is suitable for understanding the rhythm, melody, etc. of music [5]. And the proposal of MERT [6] has opened up new possibilities for music feature representation to capture the unique pitch and tonal characteristics of music by combining teacher models to provide pseudo-labels with acoustic pretraining in the style of Masked Language Modeling (MLM).

In recent years, many multimodal models for music comprehension have emerged as the research boom in large language modeling continues to rise and intensify. In terms of learning joint audiovisual representation, some studies research capturing extended temporal context in music and video modalities, providing a sophisti-

cated approach to content pairing [7]. There are also studies that attempt to combine video data for multimodal music analysis to better understand the relationship between music and audiovisual scenes [8]. In terms of learning about joint representation for speech, the proposal of MusCaps [9] mitigates semantic differences in the mir task by combining CNN and RNN network architectures, using multimodal encoders for joint processing of audio-text inputs, and pre-training with audio data to obtain representations that effectively capture and summarize musical features in the input, as well as the Song Describer Dataset (SDD) created [10], which brings new options for evaluating music and language models by using this data set to benchmark popular models.

The introduction of large language models has provided new perspectives on music understanding, especially in the combination of cross-modal learning and pre-trained models. For example, the Mu-LLAMA [11] model is based on the Llama2 framework and utilizes the MERT model for pretrained representation extraction of audio signals. This enables Mu-LLAMA to achieve excellent performance in music comprehension tasks, especially when dealing with multimodal inputs, exhibiting excellent reasoning and analyzing capabilities. In addition, the MusiLingo [12] model also combines Llama and pre-trained acoustic music representations for music subtitle generation and instruction following, thus demonstrating the wide range of applications of multimodal fusion techniques in music understanding.The domestic Qwen2-Audio [13] model has also made significant progress in the field of music understanding.Qwen2-Audio combines the Whisperlarge-v3 model with the Qwen7B large language model, and is capable of accepting inputs from a variety of audio signals and carrying out audio analysis, or responding to the corresponding text based on speech commands. This model is not limited to audio signal processing, but is also capable of combining verbal commands for music analysis and recommendation, demonstrating the great potential of AI in multimodal music understanding and interaction.

In terms of research on text-to-music generation, MusicLM [14], a pioneering technology, successfully realizes high-quality music creation based on text descriptions utilizing CLAP embedding as a text condition for music generation. This generative model is capable of generating stylistically rich and emotionally diverse music pieces based on given textual descriptions, marking the deep integration of music composition and artificial intelligence. Meanwhile, M2UGen [15] also realized the ability to generate music from multimodal inputs such as audio, text, images, and video by using large language models (LLM) based on the Whisperlarge-v3 model. These techniques not only advance the technology of music composition but also provide theoretical and technical support for cross-modal music generation in practical applications.

2.2 Music Evaluation Benchmarks

Music evaluation benchmark is an important tool for evaluating and comparing the performance of different music understanding models. With the deepening of music comprehension research, related assessment benchmarks have also been developed, mainly focusing on the construction of datasets, the definition of assessment criteria, and the construction of benchmark frameworks.

Datasets are a crucial part of music understanding and assessment. Currently, several publicly available datasets are widely used in music understanding research. Common datasets include GTZAN [16], which contains 10 different styles of music; FMA [17] (Free Music Archive), a large-scale, open-source dataset containing multiple genres of music; and MagnaTagATune [18], which provides detailed tagging information

for each music sample and is suitable for music classification and sentiment analysis among other tasks. As the diversity and complexity of datasets have increased with the application of deep learning, existing evaluation benchmarks have tended to diversify to be able to support more diverse research needs.

The evaluation of music understanding models is usually based on criteria such as classification accuracy and F1 value, but these evaluation criteria often do not fully reflect the performance of the model, especially in tasks such as sentiment analysis and style classification. Therefore, researchers have proposed more refined evaluation criteria. For example, accuracy, precision, and recall in sentiment analysis can be used to evaluate different aspects of the model separately [19]. In addition, in tasks such as style categorization, F1 scores are considered to be an effective indicator of model stability and accuracy.

In order to promote the standardization of music understanding techniques, benchmarking systems for several music understanding tasks have emerged. MuchoMusic [20] aims to assess the capabilities of multimodal audio-linguistic models for music understanding by using the MusicQA evaluation approach.MARBLE provides a comprehensive evaluation benchmark by defining a comprehensive four-tiered classification system covering acoustic, performance, score, and high-level description dimensions, enabling a comprehensive assessment of diverse downstream tasks on the model. In addition, specialized evaluation frameworks have emerged for specific languages. Muchi [21] the first open-source benchmark test for music description in Chinese, aims to evaluate the performance of multimodal LLMs in understanding and describing music. ZIQI_eval [22], targeted at large models, is a comprehensive large-scale music evaluation benchmark. It covers 10 main categories and 56 subcategories, covering a wide range of aspects such as music theory, composition, genre, instrumentation, and historical context, and is designed to assess the ability of LLMs in the field of music.

3 Method

3.1 Overview

The evaluation methodology used in the paper is shown in Figure 1:

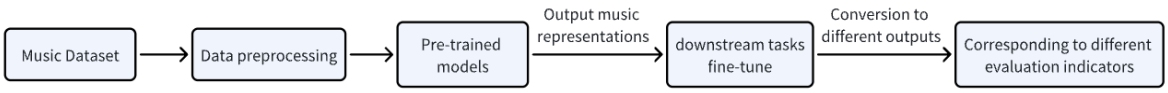

Figure 1. Evaluation process to realize the MARBLE benchmark

Six available music datasets provided in the original paper were collected for the experiments. These datasets were preprocessed, including the unification of audio formats and so on, converting part of the audio to 16kHz sampling rate, mono wav format, etc., and the data were partitioned into training, validation, and test sets according to the strategy specified in the original paper. Then, representative music pre-training models are selected and reproduced, and two versions of MERT: MERT-v0-public and MERT-v1-95M, are selected for the reproduction experiments. During the reproduction process, the network structure, parameter settings and training methods of each model are studied in depth to ensure that the hyper-parameters of the training are consistent with that of the original paper, so that the models can learn the music features accurately. Then,

the music representations are output through the pre-trained models, which are applied to the downstream tasks. The evaluation framework is built based on the unified evaluation protocol of the paper, and specific prediction heads are connected for the pre-trained model for different downstream task characteristics (classification, regression), and different training strategies are considered. Finally, the model is fine-tuned according to the different output forms of the downstream tasks, and the corresponding evaluation metrics are used to comprehensively evaluate the performance of the model in music comprehension-related tasks.

3.2 Pre-train Model for Evaluation

Due to the existence of two teacher models in the MERT framework, it increases its representation learning ability while helping the model to reduce its dependence on labeled data; in addition, it also has multiple parameter versions to choose from, combining the advantages of lightweight and efficient inference, and has excellent comprehensive performance. Therefore, two versions of MERT, MERT-v0-public and MERT-v1-95M, are chosen as the evaluation framework, and the structure of MERT is shown in Figure 2.

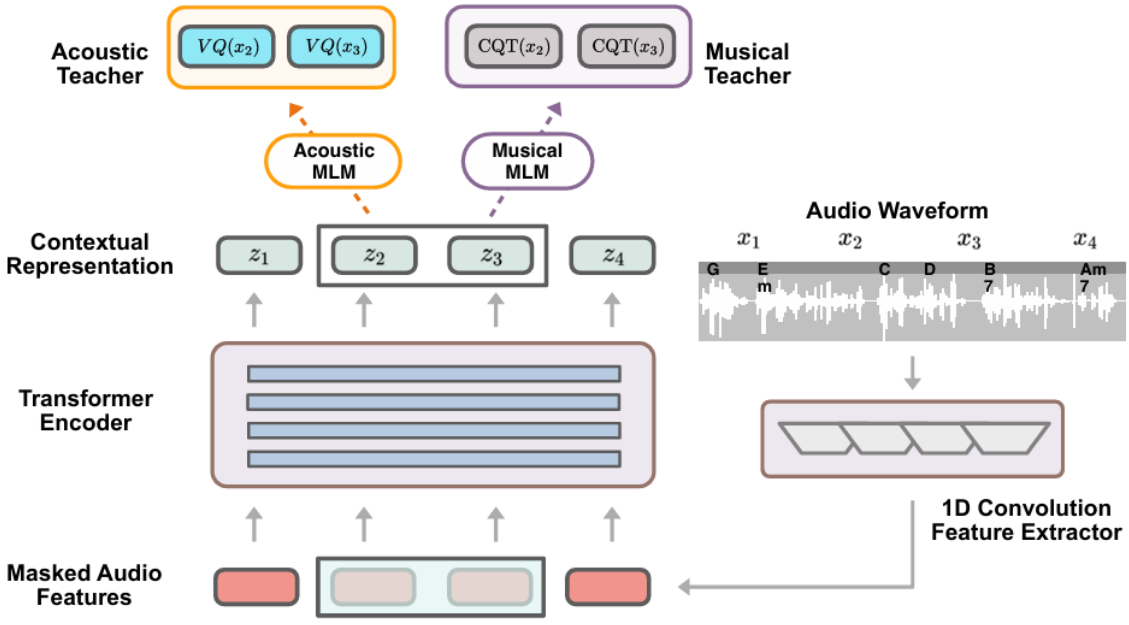

Figure 2. Illustration of the MERT Pre-training Framework.

3.3 Supported Datasets

The current supported datasets are MTG, MTT, EMO, GTZAN, GS, and VocalSet, and the following is a description of the currently supported datasets and downstream tasks:

1. MTT (magnatagatune)
    - Dataset: 25.9k 30-second audio clips totaling 170 hours with manually labeled tags.
    - Downstream task:
        - music labeling (multi-label classification, evaluation metrics are ROC - AUC macro-averaging, AP/PR - AUC).

2. MTG (MTG - jamendo)

    - Dataset: 55k audio clips, nearly 2k hours, audio can exceed 30 seconds, processed via sliding window.
    - Downstream tasks:
        - Music labeling (multi-label classification, ROC - AUC and PR - AUC/AP evaluation).
        - Genre classification (multi-tag classification, ROC and AP evaluation).
        - Emotion detection (multilabel task, ROC and AP evaluation).
        - Musical Instrument Classification (multi-label classification, ROC and AP evaluation).

3. GTZAN

    - Dataset: 30-second audio clips from 10 genres, approximately 8 hours (filtered).
    - Downstream Tasks:
        - Genre classification (multi-class classification, accuracy evaluation).
        - Beat tracking (binary classification, f_measure evaluation).

4. GS (giantsteps)

    - Dataset: 604 2-minute electronic dance tracks covering 12 major and minor pitch classes.
    - Downstream task:
        - key detection (24 classifications, weighted score evaluation).

5. EMO (emomusic)

    - Dataset: 744 45-second music clips labeled with arousal and validity scores.
    - Downstream task:
        - emotion recognition (regression task, coefficient of determination evaluation).

6. VocalSet

    - Dataset: 20 singers performing 17 techniques in 10.1 hours of audio with 3 second intervals.
    - Downstream Tasks:
        - Vocal Technique Detection (recognize singing techniques, accuracy assessment).
        - Singer Recognition (recognizing singers, accuracy assessment).

4 Implementation details

4.1 Comparing with the released source codes

This reproduction experiment is based on the open source code given in the original paper; the code is mainly divided into three parts, the data processing, the pre-training framework to be evaluated, and the

prediction header for the downstream task. In reproducing the model of the two versions of MERT-v1-95M and MERT-v0-public, we refer to the definition of the model architecture, the training parameter settings, and the data pre-processing related code in the official codebase, and select the best performing parameter settings in MARBLE for subsequent fine-tuning. code in MARBLE, and selected the best performing parameter settings for subsequent fine-tuning.

4.2 Downstreams and Training Strategies

The framework is built according to the unified evaluation protocol of the paper, and the pre-trained model is used as the backbone network to provide a generalized representation, and specific prediction heads are connected for the downstream tasks. Different types of predictor heads are designed according to the task characteristics, such as multilayer perceptron (MLP) for classification and recurrent neural network (RNN) or Transformer for sequence annotation. The evaluation process considers three training strategies: the unconstrained track allows free adjustment of model hyper-parameters and structure; the semi-constrained track requires freezing of the pre-trained backbone network; and the constrained track employs a standardized setup that restricts the hyper-parameter search space and uses a frozen model to extract features and train them in combination with structure-specific predictor heads.

5 Results and analysis

Replication of 10 tasks on 6 datasets on a single RTX4090 consumer GPU using MERT-v1-95M and MERT-v0-public as pre-training models, and the results are shown in the following Table 1:

Table 1. Evaluation results of the two models under different downstream tasks.

Datasets	**MTT**		**GS**	**GTZAN**	**EMO**		**VocalSet**		**MTG**		**MTG**		**MTG**		**MTG**	
Task	Tagging		key	Genre	Emotion		Tech	Singer	Instrument		MoodTheme		Genre		Top50	
Metrics	ROC	AP	$Acc^{Refined}$	Acc	$R2^V$	$R2^A$	Acc	Acc	ROC	AP	ROC	AP	ROC	AP	ROC	AP
MERT-V0-public	90.9	37.3	67.3	68.0	56.2	71.8	82.9	72.4	73.6	19.2	73.5	10.7	87.0	20.0	81.0	28.1
MERT-V1-95M	91.0	37.6	65.2	74.1	59.9	76.6	79.2	81.2	74.3	19.8	74.1	10.9	87.0	18.7	81.6	29.0

As can be seen from the evaluation results: different models show different advantages on each task, the overall evaluation results are similar to the original paper, and the replication experiment is valid. MERT-v0-public has a higher accuracy in the key detection task (GS Key), which indicates that it is more capable of learning pitch and tonal features; in the music tagging task (MTT Tagging), MERT-v0-public and MERT-v1-95M have similar ROC and AP values, and have comparable ability to differentiate between tags. However, overall there is still room for improvement in some tasks, for example, in the EMO Emotion task (EMO Emotion), the $R2^V$ and $R2^A$ of MERT-v1-95M is higher, but there is still a gap between it and the ideal result, probably due to the complexity of emotion recognition and the model's insufficient capturing of emotional features. Comparing the performance of different models on the same task, such as the musical instrument classification task (MTG Instrument), MERT-v1-95M has a slightly higher ROC, which may be due to its structure or pre-training learning of musical instrument features. Meanwhile, dataset characteristics also affect

model performance, e.g., the size, stylistic diversity, and labeling accuracy of the GTZAN dataset can constrain the accuracy of the model in the genre classification task (GTZAN Genre).

6 Conclusion and future work

This replication experiment successfully verifies the effectiveness and feasibility of the MARBLE benchmark in evaluating music audio representation models, and provides insights into the performance of the pre-trained model in music information retrieval tasks through multiple datasets and task experiments, which provides valuable references for music AI research. However, the experiments also show that the performance of some task models needs to be improved, and future research can improve the overall performance of the music understanding model by improving the model structure, optimizing the pre-training strategy, and expanding the dataset. With the development of the field, more innovative results are expected to promote music AI in the direction of more intelligent and accurate.

References

[1] Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Le Zhuo, Yiqi Liu, Jiawen Huang, Zeyue Tian, et al. Marble: Music audio representation benchmark for universal evaluation. *arXiv preprint arXiv:2306.10548*, 2023.

[2] Christian Bakke Vennerød, Adrian Kjærran, and Erling Stray Bugge. Long short-term memory RNN. *CoRR*, abs/2105.06756, 2021.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[4] William Grant Hatcher and Wei Yu. A survey of deep learning: Platforms, applications and emerging research trends. *IEEE Access*, 6:24411–24432, 2018.

[5] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *International Society for Music Information Retrieval Conference*, 2011.

[6] Yizhi Li, Ruibin Yuan, Ge Zhang, Yi Ma, Xingran Chen, Hanzhi Yin, Chen-Li Lin, Anton Ragni, Emmanouil Benetos, N. Gyenge, Roger B. Dannenberg, Ruibo Liu, Wenhu Chen, Gus G. Xia, Yemin Shi, Wen-Fen Huang, Yi-Ting Guo, and Jie Fu. Mert: Acoustic music understanding model with large-scale self-supervised training. *ArXiv*, abs/2306.00107, 2023.

[7] Dídac Surís, Carl Vondrick, Bryan Russell, and Justin Salamon. It's time for artistic correspondence in music and video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10554–10564, 2022.

[8] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification, 2016.

[9] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and Gyorgy Fazekas. Muscaps: Generating captions for music audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, page 1–8. IEEE, July 2021.

[10] Ilaria Manco, Benno Weck, SeungHeon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. The song describer dataset: a corpus of audio captions for music-and-language evaluation, 2023.

[11] Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. Music understanding llama: Advancing text-to-music generation with question answering and captioning, 2023.

[12] Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response, 2024.

[13] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report, 2024.

[14] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text, 2023.

[15] Shansong Liu, Atin Sakkeer Hussain, Qilong Wu, Chenshuo Sun, and Ying Shan. M^2ugen: Multi-modal music understanding and generation with the power of large language models, 2024.

[16] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[17] Kirell Benzi, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. *CoRR*, abs/1612.01840, 2016.

[18] Edith Law, Kris West, Michael Mandel, Mert Bay, and J. Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009*, pages 387–392, 2009.

[19] Ting Li. Music emotion recognition using deep convolutional neural networks. *Journal of Computational Methods in Sciences and Engineering*, 24(4-5):3063–3078, 2024.

[20] Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. Muchomusic: Evaluating music understanding in multimodal audio-language models, 2024.

[21] Zihao Wang, Shuyu Li, Tao Zhang, Qi Wang, Pengfei Yu, Jinyang Luo, Yan Liu, Ming Xi, and Kejun Zhang. Muchin: A chinese colloquial description benchmark for evaluating language models in the field of music. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7771–7779, 2024.

[22] Jiajia Li, Lu Yang, Mingni Tang, Chenchong Chenchong, Zuchao Li, Ping Wang, and Hai Zhao. The music maestro or the musically challenged, a massive music evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3246–3257, Bangkok, Thailand, August 2024.