MIREX 2025 MuCrossCover System

Mayia Chen[*] Gorta Gong[†] Wucheng Wang[‡] Lester Kong[§]
Ethan Zhao[¶]
Lyra Lab, Tencent Music Entertainment
Shenzhen, China

September 10, 2025

1 Abstract

This paper presents our submission to the cover song identification task for the 2025 edition of the Music Information Retrieval Evaluation eXchange (MIREX). This approach aims to effectively identify songs in cover scenarios. In our system, we introduce the Transformer network[1] to better capture the contextual relationship between melodies. MeanWhile, we design the model trained with multi-losses including metric learning and classification loss functions to achieve high metrics. Combining with multi-modal information, we get the submission edition which is called MuCrossCover.

2 Introduction

Cover song identification(CSI) is a crucial task in the field of Music Information Retrieval (MIR), aiming at finding the music version given a music track. Traditional audio fingerprinting algorithms can typically only identify completely identical audio versions, but perform inadequately when handling scenarios with significant variations such as live performances or adaptations. In contrast, the cover song identification task aims to learn robust feature representations that capture the essence of a song across different versions, enabling effective retrieval of cover songs that exhibit variations in tonality, tempo, structure, instrumentation, or music genre.

Currently, cover song recognition is mainly based on deep learning methods, which focus on CNN architecture, such as TPPnet[1], Bytecover2[2], WideResnet[3]. However, to solve the relationship of long sequences, CoverHunter[5] uses a conformer network to better capture local and global characteristics. Recently, some works focus on multi-modal information to promote the model performance, Such as Lyrics[6]. Based on the above ideas, we trained and obtained the final submitted system.

During the training process, we incorporated multiple loss functions for optimization, including Triplet Loss for feature distance measurement and Cross-Entropy Loss for classification. Simultaneously, we performed data augmentation on the training data, applying techniques such as Spectral Augmentation, Tempo Change, and Pitch Shift to enhance the model's generalization capability against variations.

3 Proposed Model

3.1 Model

In this system, we use Second Hand Song dataset to train our model. For feature extraction, the Constant-Q Transform (CQT) is employed with a sampling rate of 22050 Hz and a hop size of 512.

[*] Email: mayiachen@tencent.com
[†] Email: gortagong@tencent.com
[‡] Email: wuchengwang@tencent.com
[§] Email: lesterkong@tencent.com
[¶] Email: ethanzhao@tencent.com

In training stage, the CQT features are cropped with variant lengths T to enhance the robust of the model. The output dimension of the model is fixed at 512. During training, the learning rate is dynamically adjusted using a cosine annealing scheduler. In the metric learning phase, each training batch consists of multi cover groups, with each group encompassing 4 different versions. The Adam optimizer is utilized with an initial learning rate set to 0.001.

3.2 Retrieval

Our retrieval system utilizes Maximum Inner Product Search (MIPS). Given that our model produces L2-normalized embeddings, this MIPS operation is mathematically equivalent to performing a maximum cosine similarity search.

3.3 System Implementation

Our implementation is based on PyTorch, and simplifies deployment through unified export to the ONNX format. By leveraging optimized ONNX runtime libraries natively supported by target platforms, it eliminates the cumbersome work of separately adapting PyTorch models for each specific environment.

3.4 References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.

[2] Z. Yu, X. Xu, X. Chen, and D. Yang, “Temporal pyramid pooling convolutional neural network for cover song identification,” in Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019, pp. 4846–4852.

[3] X. Du, K. Chen, Z. Wang, B. Zhu, and Z. Ma, “Bytecover2: Towards dimensionality reduction of latent embedding for efficient cover song identification,” in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 616–620.

[4] S. Hu, B. Zhang, J. Lu, Y. Jiang, W. Wang, L. Kong, W. Zhao, and T. Jiang, ”WideResNet with joint representation learning and data augmentation for cover song identification”. In Proc. of the Conf. of the Int. Speech Association (INTERSPEECH), 2022, pp. 4187–4191.

[5] F. Liu, D. Tuo, Y. Xu, and X. Han. CoverHunter: cover song identification with refined attention and alignments. In Proc. of the IEEE Int. Conf. on Multimedia and Expo (ICME), pp. 1080–1085, 2023.

[6] M. Balluff, P. Mandl, and C. Wolff, “Innovations in cover song detection: A lyrics-based approach,” arXiv preprint arXiv:2406.04384, 2024.