# Introduction and Literature Survey on Music Similarity and Copy Detection

## Introduction

Music has always been more than entertainment. It has patterns, rhythms, and melodies that make us feel emotions and connect across cultures. With the growth of streaming platforms and social media, there are now millions of tracks easily accessible to anyone. This creates both opportunities and challenges. One important challenge is to identify when songs are similar or when a new release may be copying an old one. This is not only important for copyright and legal protection but also for organizing mus...

## Literature Survey

From the first set of twelve papers I studied, I saw a wide range of approaches. One paper focused on using lyrics transcription through ASR and then matching the text while also checking audio features. This worked well for songs where lyrics are important but less for instrumental covers. Another paper created a framework to test multiple similarity measures together. This was helpful because it showed not only which methods performed well but also that no single method works best in every situation.  
  
A very interesting direction came from a paper that used self-supervised learning for music similarity. The model was trained without labels using contrastive learning and triplet loss. The idea was that if two clips are from the same song, they should be close in the embedding space, and if not, they should be far apart. This method is powerful because it reduces the need for annotated datasets, which are hard to get for music. Another work looked at artist similarity not only from audio but also using ...  
  
Later works made a big shift by introducing transformer-based methods. One paper showed how transformers could be trained for cover song identification, achieving better performance than CNNs because they capture long-term dependencies such as melody and rhythm. Another line of work introduced the ByteCover series, with ByteCover3 being the improved version. It used embeddings along with a special local alignment loss to match songs more precisely. This gave excellent results on datasets like SHS100K. S...  
  
Apart from deep models, I came across a detailed survey paper that explained how the field has grown. It started from MFCC and DTW methods, moved to chroma and HPCP features, and then reached CNNs, transformers, and multimodal fusion. This survey was valuable for me because it connected my own implementation path with the bigger picture in research. In addition, some papers looked at fast copyright detection using perceptual hashing. The HADES system, for example, used robust hash functions on audio sign...  
  
I also saw papers that were closer to audio forensics than music similarity. One of them used keypoints on Mel-spectrograms, similar to how SIFT is used in images, to detect copy-move forgeries inside audio. Another used signal decomposition with VMD and MRMR for localizing manipulated sections. While these are not about cover detection, they share the same goal of identifying copied or altered parts in audio.  
  
When I expanded my survey to more recent thirty papers, I noticed that multimodal approaches became much more common. Researchers began combining audio embeddings with lyrics and metadata. This makes sense in platforms like YouTube, where titles, tags, and descriptions can reveal relationships that audio alone may not. New datasets like LyricCovers 2.0 were built to support this. There were also papers that directly tested lyric semantic similarity using sentence transformers and fused it with audio si...  
  
I also read about systems for query-by-humming (QbH), which rely almost fully on melody. Semi-supervised learning helped improve QbH models, making them more robust when people hum with different tempos or styles. In contrast, the transformer and conformer-based works like CoverHunter did not focus only on melody but also modeled temporal rhythms and spectral timbre, making them powerful across all aspects of music similarity.

## My Work and Implementation

After reading these papers, I decided to follow a similar journey in my project. I started simple by using MFCCs. I took a few songs, extracted MFCC features with librosa, and compared them using distance measures. This was useful to learn how timbre can be represented numerically. However, I quickly noticed that MFCCs were not enough when the songs had different instruments but the same melody.  
  
To improve, I implemented Dynamic Time Warping (DTW). DTW allowed me to align two feature sequences even if ...

## Observations and Reflections

By implementing MFCCs, I understood why they are so widely used as a first step — they are simple, efficient, and describe timbre. With DTW, I learned how important alignment is for music, because songs may have the same melody but played at different speeds. With CNNs, I saw how deep learning can automatically extract features that are difficult to design by hand. Comparing these results with what I read in research papers made me appreciate the progress in this field.  
  
I also realized the importance ...

## Conclusion

From my reading and experiments, I see that music similarity and copy detection is a field that has evolved step by step. It started with handcrafted features like MFCCs and DTW, moved to CNNs, and now is advancing with transformers and multimodal approaches. Some methods focus strongly on melody, others on timbre, and some on rhythm. The most advanced works today combine all of these and even add lyrics and metadata.  
  
My own project followed a similar path. I started with MFCCs and DTW to build unders...

## References

1. Vaglio, A. et al. Cover Detection with Lyrics Transcription. ISMIR 2021.

2. Bodo, R. et al. A Framework for Music Similarity and Cover Song Identification. CMMR 2021.

3. Thomé, C. et al. Musical Audio Similarity with Self-supervised Features. arXiv 2022.

4. Korzeniowski, F. et al. Artist Similarity for Everyone: A GNN Approach. Transactions of ISMIR 2022.

5. Zeng, T. et al. Training Audio Transformers for CSI. EURASIP Journal 2023.

6. Du, X. et al. ByteCover3: Embedding and Local Alignment. arXiv 2023.

7. Liu, F. et al. CoverHunter: Refined Attention and Alignment. arXiv 2023.

8. Zheng, Y. Cover Song Identification Technologies: A Survey. ACM Computing Surveys 2023.

9. Ansori, M. et al. HADES: Hash-Based Audio Copy Detection. IEEE Transactions 2023.

10. Ustubioglu, B. et al. Audio Copy-Move Forgery Detection. Expert Systems with Applications 2023.

11. Hachmeier, S. & Jäschke. Cover Song Detection via Metadata Fusion. NLP4MusA 2024.

12. Balluff, M. et al. LyricCovers 2.0: Dataset for Cover Song Analysis. IADIS 2024.

13. Balluff, M. et al. Innovations in Lyrics-Based Cover Detection. arXiv 2024.

14. Amatov, A. et al. Semi-Supervised QbH and Similarity Tasks. ISMIR 2023.

15. X-COVER: ASR-Lyric Fusion for Version Identification. Zenodo 2024.

16. Peng, K. et al. Robust Copy-Move Detection with VMD-MRMR. Elsevier 2024.

17. Nasr, M. et al. Survey of Audio Forgery Detection. Springer 2025.

18. Ellis, D. Chroma-based Cover Song Identification with DTW. ISMIR 2007.

19. Serra, J. et al. Cross Recurrence Plots and Cover Detection. IEEE TASLP 2009.

20. Bello, J. et al. Audio Similarity and Alignment with HPCP Features. ISMIR 2010.

21. Humphrey, E. et al. Deep Learning for MIR: A Survey. ISMIR 2013.

22. Dieleman, S. & Schrauwen, B. CNNs on Spectrograms for Music. ISMIR 2014.

23. Choi, K. et al. Convolutional Recurrent Models for Music Classification. ICASSP 2017.

24. Won, M. et al. Evaluation of Transformers vs CNN for MIR. arXiv 2022.

25. Bertin-Mahieux, T. et al. The Million Song Dataset. ISMIR 2011.

26. Defferrard, M. et al. FMA: A Dataset for Music Analysis. ISMIR 2017.

27. Salamon, J. et al. Deep Convolutional Neural Networks for Music Auto-Tagging. IEEE TASLP 2017.

28. Doras, G. et al. Contrastive Learning for Music Similarity. ISMIR 2021.