# Introduction and Literature Survey on Music Similarity and Copy Detection

## Introduction

Music is not only a form of art but also a language that carries patterns, emotions, and structures. In today’s digital world, millions of songs are being produced and shared every day, and with that comes the question of how to detect similarity or copying between them. This is important not only for copyright issues but also for recommendation systems, music organization, and research in music information retrieval. When I first thought about this project, I wanted to see how similar songs can be detected based on their sound, lyrics, or structure. At the beginning, I started experimenting on my laptop with smaller datasets. Later, I realized that larger datasets would give more reliable results, so I shifted to using the FMA dataset, which contains more songs compared to the MTA dataset. To handle audio files, especially compressed MP3s, I also had to use ffmpeg for uncompression and librosa for feature extraction and signal analysis.  
  
My journey in implementation also followed the same path that the field has seen in research: starting with basic hand-crafted features and distance measures, then moving towards advanced deep learning methods. At first, I tried with MFCCs, which are among the most classical features to describe timbre. I then explored Dynamic Time Warping (DTW) to align sequences and check similarity even when timing is different. As I went deeper, I learned that many recent research papers have moved towards deep learning methods such as CNNs, transformers, and multimodal approaches that mix lyrics, audio, and metadata. Reading these papers gave me both inspiration and direction for my own project.

## Literature Survey

When I went through the earlier set of twelve papers, I noticed that each one approached the problem of music similarity and cover detection in slightly different ways. One of the papers worked on combining lyrics with audio. The idea was that many covers keep the same lyrics even if the instruments and voices change, so using automatic speech recognition (ASR) to convert audio into text and then matching it with audio features gave them better results, especially when the lyrics were clearly available. Another paper created a whole benchmarking framework where they tested many different methods together, including both hand-crafted features and machine learning models. This helped to show that there is no single perfect method, but some techniques worked better depending on the dataset and conditions.  
  
I also found a paper that tried self-supervised learning for music similarity. Instead of relying on labeled data, they trained a network using contrastive learning and triplet loss. This was interesting because music datasets are huge, and labeling is hard. Their model was able to learn useful representations directly from audio and then compare songs at the segment level. A related work looked at artist similarity using graph neural networks, mixing both graph connections (like related artists, playlists, and co-listens) with content features from audio. This showed that similarity can also be understood in terms of context, not only in terms of sound.  
  
More recent papers focused on using transformers for audio. One of them explained how transformers could be trained for cover song identification and found that these models often did better than CNNs because they can capture long-term patterns like melody and rhythm. Another group introduced the ByteCover family, especially ByteCover3, which used embeddings and local alignment to reach very strong performance on benchmarks like SHS100K. Along with this, CoverHunter took things further by refining attention and chunk alignment, showing even higher accuracy. These models stood out because they tried to model not only the timbre but also the structure of a song, making them good at finding covers that change instruments or arrangements but keep the core tune.  
  
I also read surveys and resource papers. One survey gave a clear overview of how the field developed, from classical MFCC + DTW methods up to deep learning and multimodal approaches. Another line of work looked at hashing, such as the HADES system, which focused on copyright detection. Instead of deep networks, it relied on perceptual hashing that can quickly check whether an audio clip has been copied or altered. This method is very efficient and useful when the main aim is to protect rights rather than find artistic similarity. There were also forensic-style papers, like one that used keypoints on Mel-spectrograms, similar to how SIFT works in images, to detect copy-move forgeries within audio. This was more about manipulation detection than musical similarity, but it still showed how spectral features can reveal copied sections.  
  
When I extended my reading into the recent thirty papers, I noticed a clear trend towards combining different modalities and using more powerful models. For example, some new works combined lyrics, metadata, and audio embeddings to detect covers on platforms like YouTube, where user tags and descriptions give additional clues. Datasets like LyricCovers 2.0 were released to support this line of research. Others went deeper into lyric-based detection by using sentence-transformer models that can understand the meaning of lyrics and combine that with audio features. This gave an advantage in situations where the lyrics are shared but the music is very different.  
  
Another direction was in query-by-humming (QbH) and melody-based systems. These focus almost entirely on melody, which is often the most recognizable part of a song. Semi-supervised learning was used in one paper to make QbH systems better, and this again showed how melody-based similarity can be powerful. On the other side, transformer and conformer-based systems like CoverHunter not only focused on melody but also captured rhythm and timbre through temporal attention and chunk-wise alignment. These gave some of the best accuracy results reported in recent years.  
  
In terms of timbre-focused methods, perceptual hashing methods stood out. These methods ignore melody and rhythm and focus on spectral fingerprints, making them very good for fast detection. They may not be able to find creative covers where the timbre changes but the tune stays the same, but for duplicate or copyright protection they are almost unbeatable. Forgery detection papers like the one that used VMD-MRMR or decomposition methods showed how signal processing can still compete in cases where localization of copied sections is needed.  
  
Looking at all these works together, I can see a clear evolution. The older methods like MFCC with DTW were simple but effective, especially when datasets were small. They taught me the basics of how to align sequences and compare them. As datasets got larger and more diverse, the need for deep learning became stronger. CNNs on spectrograms were a natural step, and then transformers pushed this further by modeling long-term structures. Alongside this, the move towards multimodal approaches with lyrics and metadata shows that in the real world, similarity is not only about sound but also about context and information around the music.  
  
In my own implementation, I followed a similar path. I started with MFCCs and DTW because they are easier to code and understand, and they gave me a sense of how similarity can be measured. But soon I realized their limits, especially when timing or instruments change. That’s why I moved towards CNNs with spectrograms, and I am now looking at more advanced methods inspired by works like CoverHunter and ByteCover. Switching to Google Colab gave me the power to train larger models, and using the FMA dataset provided me with a richer collection of songs compared to MTA. Step by step, I can see how my project mirrors the way research in this field has also evolved.