# REAL-WORLD MUSIC PLAGIARISM DETECTION WITH MUSIC SEGMENT TRANSCRIPTION SYSTEM

Seonghyeon Go
Mippia Inc.
E-mail: gsh@mippia.com

***Abstract*—As a result of continuous advances in Music Information Retrieval (MIR) technology, generating and distributing music has become more diverse and accessible. In this context, interest in music intellectual property protection is increasing to safeguard individual music copyrights. In this work, we propose a system for detecting music plagiarism by combining various MIR technologies. We developed a music segment transcription system that extracts musically meaningful segments from audio recordings to detect plagiarism across different musical formats. With this system, we compute similarity scores based on multiple musical features that can be evaluated through comprehensive musical analysis. Our approach demonstrated promising results in music plagiarism detection experiments, and the proposed method can be applied to real-world music scenarios. We also collected a Similar Music Pair (SMP) dataset for musical similarity research using real-world cases. The dataset are publicly available.**[1]

## I. Introduction

Music plagiarism is one of the most important copyright issues in society. The unauthorized copying of musical elements can have serious legal and economic consequences [1]. Contrary to the definition of a word, the commonly used word ”music plagiarism” can be controversial enough even if it is not intentional by the musician. Therefore, technology for detecting plagiarism can be useful for both original composers and alleged plagiarists. With the advancement of AI music generation, creating and distributing music has become more accessible, making plagiarism detection important.

Research on defining musical similarity and detecting music plagiarism has been conducted widely [2][3]. However, applying these studies to real audio data faces several challenges. Most plagiarism detection research relies on MusicXML or MIDI formats, while commercial music exists as raw audio, requiring transcription. Also, many studies assume melodically similar music is plagiarized, but this differs significantly from real-world cases. And real plagiarism cases are complex [1], potentially including vocals, varying in length, or containing brief plagiarized segments within longer tracks. A proper model needs to identify musically meaningful segments and detect plagiarism within them.

To address these issues, we propose transcribing raw audio into musical representations to organize essential musical features. Our goal is extracting musically meaningful and quantized data for plagiarism detection. Although similar ideas exist [4], we focus on creating structured segment optimized for plagiarism detection by combining various music information retrieval techniques. Based on these quantized data, we explain how to detect plagiarized music using similarity metrics. Finally, we construct a Similar Music Pair (SMP) dataset containing metadata of similar music pairs with timestamps of similar segments.

[1]https://github.com/Mippia/smp_dataset

## II. Related Works

### A. Music Transcription

Music transcription extracts note information from raw audio, typically producing MIDI representations. This task has been studied across various genres and instruments [5] or metadata like lyrics [6]. Beyond MIDI, there is growing interest in transcribing audio into music-score-like representations [7]. This approach allows transcription of complete musical progressions with temporal components, such as measures.

We propose segment transcription that incorporates music structure analysis and introduces metric-based self-similarity to analyze and transcribe music segments while adding musical information.

This approach allows detailed and musically meaningful results beyond calculating similarity. For instance, we could pinpoint that music A’s first chorus segment (e.g., 00:35–00:43, 16th–19th bar) corresponds to music B’s second chorus segment (e.g., 01:42–01:51, 44th–47th bar). We define this musical unit as a ’Segment’. This methodology enables detection of similar or plagiarized segments across larger datasets.

To perform music segment transcription, we combined MIR technologies including music source separation, beat-tracking, chord recognition to obtain necessary metadata and construct better structural representations for each segment.

### B. Music Plagiarism Detection

Research on music plagiarism analysis employs various methodologies, including CNN-based approaches [2], bipartite graph-based methods [3], NLP-based methods using tokenization [8], and audio fingerprinting-based methods [9]. However, most approaches focus on MIDI or MusicXML data, with limited methodologies using raw audio data [10]. But in

979-8-3315-7206-8/25/$31.00 

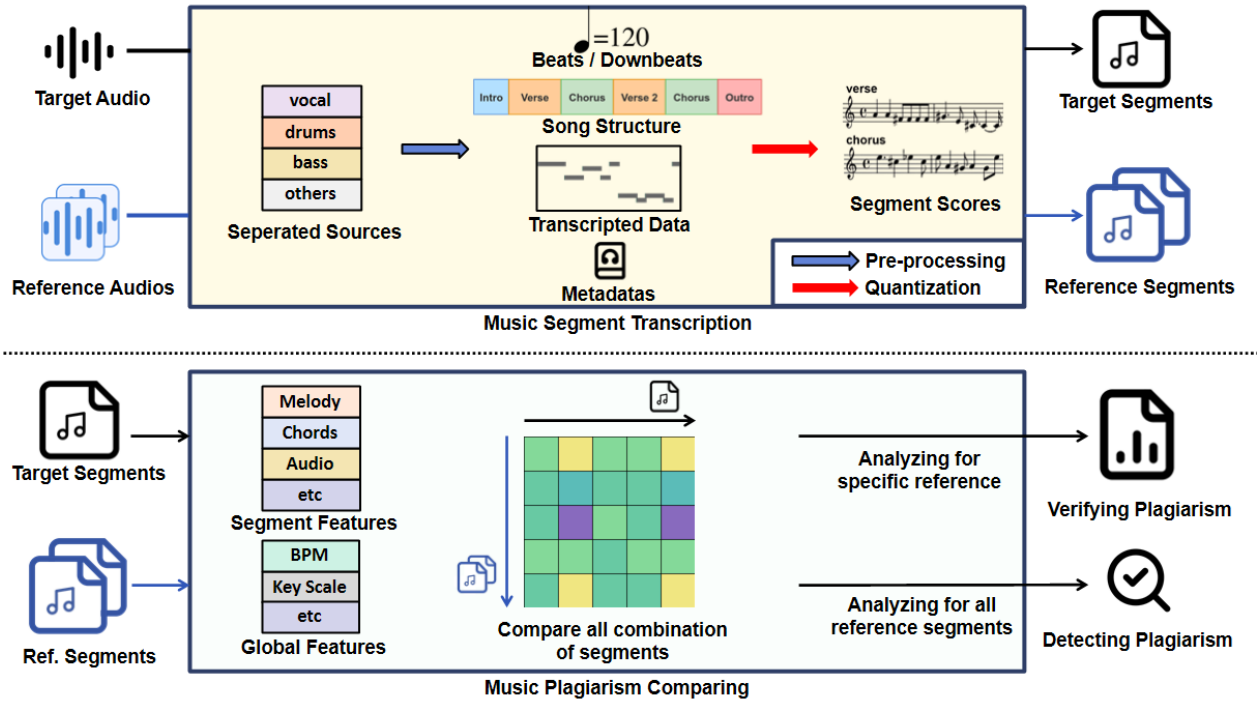


Fig. 1. Overall structure of music plagiarism detection system

applying plagiarism detection to real-world scenarios, using raw audio data is essential.

Cover Song Identification (CSI) can be considered a similar task to plagiarism detection, as it involves retrieving cover versions of a query song from a music dataset. Classically, there is a melody MIDI-based methodology [11] or a methodology for interpreting and comparing a sound source as a sequence [12]. Recently, the trend of studying CSI based on deep learning is increasing. Various methods using CNN have been proposed [13]–[15]. In addition to this, models using the conformer structure with music chunk [16] have achieved SOTA performance. However, CSI methods typically perform song-to-song comparisons to determine overall similarity between entire musical works. In contrast, plagiarism detection requires more fine-grained analysis to identify specific passages where copying occurs, necessitating segment-level comparisons that can pinpoint exactly which parts of songs are similar and potentially plagiarized.

A key challenge in plagiarism detection is defining MIDI or audio similarity. Some approaches use shape similarity [17], while others tokenize notes and apply sequence-based deep learning models [18]. Other approaches train models to learn features similar to plagiarism cases, using embedding distance as metrics [2]. We propose a segment-based methodology that decomposes songs into meaningful musical segments and performs individual comparisons between segments. This approach enables detection of partial plagiarism where only specific sections are borrowed and provides precise localization of similar passages. We combine various metrics after music segment transcription, focusing on shape-based similarities and musical feature similarities.

## III. Proposed System

The overall structure is illustrated in Figure 1. The target audio is processed through the music segment transcription system. After preprocessing and quantization, the audio is analyzed in distinct segments. Each segment contains musical information such as melody, chords, instruments, etc. These segments are then compared for similarity with other preprocessed reference segments, enabling plagiarism detection. In the following sections, we provide a detailed explanation of how the music segment transcription process is carried out and how the similarity computation is performed.

### A. Music Segment Transcription

We first preprocess music to find necessary components for quantization, and then quantize music into segments. Necessary information from segments is described in Table I, with a music plagiarism detection scenario. The preprocessing required for this task includes source separation, transcription results of each separated source, chord recognition results, music structure information, tempo, downbeat, time signature, and optionally, lyric information or various metadata of the music.

To transcribe music segments, we first identify the bpm and downbeats of the waveform $w$ and quantized to a arithmetic sequence, denoted as $db_w = \{dbt_1, dbt_2, ..., dbt_n\} \in R^n$. We assume that the segment start points align with $db_w$. We then identify the music structure to locate the points of structural

TABLE I
NECESSARY INFORMATION FROM SEGMENTS IN MUSIC PLAGIARISM DETECTION SCENARIOS

| **Scenario:** We have music A, and a dataset including music B. We want to determine that "The chorus vocal melody in music A from 30-38 seconds is plagiarized from music B's chorus vocal melody at 40-47 seconds." | |
|---|---|
| **Critical Questions** | **Necessary Information** |
| 1. What is the basis for selecting *30s* as the segment starting point in music A? How about B? | Music structure analysis, downbeat, BPM |
| 2. What is the basis for selecting *38s* as the segment ending point in music A? How about B? | Downbeat, rhythm, BPM |
| 3. How is the length discrepancy (8s vs 7s) addressed when comparing segments? | Quantized note information |
| 4. What evidence supports that the *chorus* aspect is plagiarized rather than other musical structures? | Music structure information for each segment |
| 5. What evidence supports that specifically the *vocal* component is plagiarized? | Instrument-specific transcription & similarity |
| 6. What evidence supports that the *melody* aspect is plagiarized rather than other musical elements? | Similarity for each musical element |
| 7. What evidence supports that *music B* is the source rather than other pieces? | Similarities from all reference segments |

change. These music structure boundaries are represented as $ss_w = \{ss_1, ss_2, ..., ss_m\} \in R^m$, which is quantized to be a subset of $db_w$. Since the length of each music structure is not always consistent (e.g., 4 bars or 8 bars), and even segments with the same structural role may exhibit distinct patterns (e.g., verse1 and verse2), we perform additional segment clustering to properly identify and group these varying musical elements.

We use a self-similarity-based clustering algorithm to find each start point of segments, identifying and organizing recurring core patterns that frequently appear in music. A similarity matrix is constructed by computing the similarity between all available segment start points from $db_w$. Then, we apply hierarchical clustering using Ward's method [19]. Ward's method minimizes the variance within clusters by recursively merging the pair of clusters that leads to the minimum increase in total within-cluster variance. The number of clusters is determined proportionally to the size of the distance matrix, with an appropriate threshold. After clustering, the most frequently occurring patterns in music are treated as the most significant. Note that the "distance" can be defined in any methods. Used distance between each segment is discussed in Section 3.2, which has an inverse relationship with similarity.

For each boundary identified in the music structure, we extract a potential starting point and refine these initial points by ensuring they align with musical phrases, with more segments aligned in clusters that have high priority, typically occurring at 4-bar intervals. This process yields a set of refined starting points $SP_w = \{sp_1, sp_2, ..., sp_k\}$ that accurately delineate the beginning of each significant musical segment, ensuring that our segmentation respects the inherent structural organization of the music. This approach tends to create compact segments and is effective for our segment analysis as it groups similar musical patterns while maintaining boundaries from music structure information. These starting points mark where each music segment begins and help us quantize the music's structure. For each starting point, we define a segment as the interval spanning a fixed length, typically 4 bars for this work.

We obtain transcription data for each instrument, $N_{(i,\mathcal{I})} = (p_{(i,\mathcal{I})}, t_{(i,\mathcal{I})}, d_{(i,\mathcal{I})}, v_{(i,\mathcal{I})})$ where $p$ denotes pitch, $t$ denotes onset time, $d$ denotes duration, $v$ denotes velocity in MIDI, and $\mathcal{I}$ is the set of instrument types. We obtain the start time and duration in seconds. With BPM and downbeat information, we can determine which $sp_i$ these notes correspond to and quantize their positions. Let $QN_{(i,\mathcal{I})} = (p_{(i,\mathcal{I})}, b_{(i,\mathcal{I})}, pos_{(i,\mathcal{I})}, qd_{(i,\mathcal{I})}, v_{(i,\mathcal{I})}) \in Z^5$ represent this quantized note information, where $b$ is bar number in music score, $pos$ is quantized onset in corresponding bar, and $qd$ is quantized duration. For example, we can state "The E5 vocal note exists as a quarter note on the second beat of 14th bar with a velocity of 100, and bars 12–15 comprise the verse 2 segment." Through this process, we analyze and organize the given musical data similar to sheet music format. This structured data is directly used in similarity comparison tasks.

### B. Music Plagiarism Detection

Our similarity calculation incorporates multiple musical aspects. Each aspect is structured in a algorithmic manner, allowing consideration of melody and chords based on cases found in various music plagiarism scenarios [1]:

- **Pattern Similarity** $p$: The chromagram-based intersection similarity.
- **Musical Complexity** $m$: Count of used pitches, weighted to $p$ to avoid overly simple similarity cases, such as rap.
- **Rhythmic Correlation** $r$: Jaccard similarity with quantized onset timing.
- **BPM Difference Ratio** $b$: Linear scaling of tempo relationship.
- **Chord Similarity** $c$:
  - **Roman numeral similarity** $R_n$: Functional harmony-based comparison [20]
  - **Chord quality similarity** $Q$: Major/minor and seventh chord relationships

$$c = w_R \cdot R_n + w_Q \cdot Q \tag{1}$$

TABLE II
SAMPLE ENTRIES FROM OUR SIMILAR MUSIC PAIR DATASET

| Original Title | Comparison Title | Relation | Original Time | Comparison Time | Pair # |
|---|---|---|---|---|---|
| Electric | Electric (remix) | Remake | [8, 16, 70, 78] | [7, 15, 82, 90] | 21 |
| Shiki no uta | Bul-ggot | Plagiarism Case | [73, 82, 134, ...] | [85, 93, 136, ...] | 29 |
| Volevo un gatto nero | Black Cat Nero | Remake | [25, 33, 59,...] | [68, 76, ...] | 31 |
| No scrubs | Shape of you | Plagiarism Case | [31, 41, 72, ...] | [15, 96] | 64 |

$$\text{similarity} = (\alpha + m \times p) \times (\beta \cdot \max(r, p)) \times b^{\delta} + \gamma \cdot c \quad (2)$$

Equation 2 is one example of our similarity metric used in experiments. The parameters $\alpha$, $\beta$, $\gamma$, $\delta$ and weights $W_R$, $W_Q$ balance the contribution of each musical component.

## IV. DATASETS

For our experiments, we compiled a SMP dataset, which is a comprehensive dataset of music piece pairs for plagiarism detection evaluation. The SMP dataset contains 70 pairs of original and comparison music pieces, each with relevant metadatas and segment time where similar part starts. Table II presents a sample of the SMP dataset. The SMP dataset was carefully curated to include a diverse range of music genres, release periods, and similarity types. It encompasses well-known plagiarism cases, legally disputed works, pieces with acknowledged influence, and even some pairs with coincidental similarities. This diversity allows for a comprehensive evaluation of our plagiarism detection approach across various scenarios encountered in real-world music copyright dispute. Additionally, we used the Covers80 dataset for experiments, which consists of 80 groups of cover songs, with each group containing multiple versions of the same original song performed by different artists.

## V. EXPERIMENTS

### A. Experimental Setting

Given the availability of various MIR models, we were able to implement a comprehensive system for the music segment transcription process. Specifically, we used Demucs [21] for source separation, the all-in-one model [22] for structural analysis, Beat-Transformer [23] for downbeat tracking, AST[24] for vocal transcription, SheetSage [25] for melody transcription, and Harmony Transformer [26] for chord transcription.

To evaluate plagiarism music detection, we conducted experiments using the SMP dataset and covers80 dataset. We calculated the average ranking and accuracy to determine how plagiarized music typically ranks. We conducted this experiment with parameter settings to compare each similarity metric's characteristics. With default setting ($\alpha = 1, \beta = 1, \gamma = 1, \delta = 0.5, w_R = 0.85, w_Q = 0.15$) for all experiments, we tested chord only ($\beta = 0$), MIDI only ($\gamma = 0$), rhythm only (similarity$= r$), and pattern only (similarity$= p$) cases for plagiarism music detection.

### B. Evaluation Methodology

We evaluate our plagiarism detection system using two complementary metrics.

**Segment-level Evaluation:** To further validate our segment-based approach, we conducted plagiarism segment detection experiments using the Covers80 dataset. Unlike traditional cover song identification methods that perform song-to-song comparisons, our approach conducts segment-level comparisons, enabling more fine-grained analysis of musical similarities within songs and detection of potential plagiarism at the segment level.

We employ a retrieval-based evaluation where query segments are matched against a database of source segments. Performance is measured using Precision@K, which calculates the proportion of correct matches within the top-K retrieved segments. A correct match is defined as a segment pair from the same cover song group, indicating a potential plagiarism relationship.

**Song-level Evaluation:** Since there are multiple segment pair similarities in one track pair, we computed plagiarism rates by summing the 20 highest similarity scores for each pair and ranking them in descending order. Following established practices in music similarity research [3], we use three primary metrics: *Top Average Index*, representing the average rank of the correct plagiarized song in our retrieval results, with a maximum penalty for undetected cases; *Top-1 Accuracy*, measuring the proportion of queries where the correct match appears as the highest-ranked result; and *Top-5 Accuracy*, providing a more lenient evaluation by considering whether the correct match appears within the top-5 results.

## VI. RESULTS

For song-level detection, we aggregate segment-level similarities using a weighted scoring approach. For each query song, we extract multiple segments and retrieve the top-5 most similar segments from our database for each query segment. The final song similarity score is computed as a weighted combination of these segment-level similarities, where weights are determined by the individual segment similarity scores.

### A. Plagiarism Segment Detection Results

The results in Table III demonstrate the effectiveness of our segment-based approach for plagiarism detection. Precision at top-100 indicates that our similarity metric successfully identifies plagiarism relationships at the segment level.

The segment-based approach offers several advantages over traditional whole-song comparison methods:
(1) This provides more detailed analysis by identifying specific musical passages that contribute to plagiarism relationships.
(2) It enables partial matching where only certain segments of songs are similar.

TABLE III
PLAGIARISM SEGMENT DETECTION RESULTS

| Evaluation Metric | Top-100 | Top-1000 |
|---|---|---|
| Precision@K | 98.00% | 51.80% |
| Correct Retrievals | 98/100 | 518/1000 |
| Metric Range | [73.22, 98.63] | [58.24, 98.63] |

TABLE IV
EXAMPLES OF SEGMENT-LEVEL DETECTION RESULTS FROM TOP-100 RETRIEVALS

| | |
|---|---|
| **Correct Matches** | |
| **Example 1** | Score: 98.63 |
| Query | Blue Collar Man (Styx) at 118.1s |
| Source | Blue Collar Man (REO Speedwagon) at 143.6s |
| **Example 2** | Score: 95.08 |
| Query | September Gurls (Big Star) at 9.2s |
| Source | September Gurls (Bangles) at 8.0s |
| **Incorrect Matches** | |
| **Example 3** | Score: 74.84 |
| Query | Gold Dust Woman (Sheryl Crow) at 184.2s |
| Source | Tomorrow Never Knows (Beatles) at 110.2s |
| **Example 4** | Score: 73.41 |
| Query | Hush (Milli Vanilli) at 157.4s |
| Source | Night Time Is The Right Time (Aretha Franklin) at 95.8s |

(3) This offers better interpretability by highlighting which musical elements drive the similarity scores, making it particularly valuable for plagiarism detection applications.

### B. Plagiarism Music Detection Results

The experimental results using segmentation data are presented in Table V. Results demonstrate that our system can identify plagiarism pairs using only audio data. Furthermore, they validate that incorporating various musical knowledge approaches enhances the detection of musical similarities.

Nevertheless, at the current stage, these metrics may not be considered sufficiently reliable, especially plagiarism music detection task. We present several observations regarding failure cases below:

**Lack of end-to-end segmentation model.** In this study, we performed the segmentation task by combining existing MIR systems. The integration of these systems introduces instability, as each component creates bottlenecks.

**Instability of segment similarity.** Segment musical similarity is a challenging concept to perfect at present. We expect this can be improved through future research, using both transparent and learned approaches.

**Bridging segment-level and music unit-level metrics.** The connection between segment-level similarity metrics and music unit metrics (such as cover song identification) requires further investigation. Current approaches may miss important relationships between local musical patterns and broader segment similarities, potentially affecting the overall detection accuracy.

TABLE V
PLAGIARISM DETECTION RESULTS WITH SIMILARITY CONDITIONS

| Conditions | Avg. Index | Top-1 Acc. | Top-5 Acc. |
|---|---|---|---|
| Pattern only | 13.23 | 0.2357 | 0.3143 |
| Rhythm only | 13.29 | 0.1429 | 0.2786 |
| Chord only | 12.73 | 0.1500 | 0.2929 |
| MIDI only | 11.16 | 0.2429 | 0.4071 |
| **All (SMP)** | **7.31** | **0.3786** | **0.6286** |
| **All (Covers80)** | **13.46** | **0.475** | **0.575** |

TABLE VI
PERFORMANCE WITH REAL PLAGIARISM PAIRS FROM SMP DATASET

| Models | Avg. Index | Top-1 Acc. | Top-5 Acc. | Loss |
|---|---|---|---|---|
| MERT | 4.83 | 0.333 | 0.833 | 0.0639 |
| Music2Vec | 5.00 | 0.333 | 0.667 | 0.0432 |

### C. Ablation Studies

For the SMP dataset, since the starting points of actually similar segments are provided, we can construct a Siamese network[27] for similarity metrics by performing segment transcription. We conducted a simple siamese network by obtaining embeddings from MERT[28] and Music2Vec[29]. We perform music detection task with models, with 55 training pairs and 15 test pairs.

The results are presented in Table VI. This is showing that this approach has sufficient research potential, but they also indicate that larger-scale data and better methodologies are still remain as future work.

## VII. FUTURE WORK AND CONCLUSION

In this paper, we proposed a music plagiarism detection system that effectively handles real-world audio data through music segment transcription. By combining various MIR technologies and introducing a segment-based analysis framework, our approach demonstrates promising performance while maintaining practical applicability to commercial music. This shows that segment-based transcription can effectively bridge the gap between theoretical plagiarism detection research and real-world applications.

Several research directions can enhance this system further. Each MIR component (source separation, transcription, beat tracking) could be improved for more stable segment transcription. The development of an end-to-end approach for segment transcription represents a promising direction that could improve both efficiency and accuracy. The structured segment data format could be extended to other MIR tasks beyond plagiarism detection, including cover song detection and music generation. Future research on deep learning architectures trained on segment-level data could capture more detailed relationships between musical pieces, strengthening the system's capability to handle real-world audio data and contributing to music copyright protection and analysis.

## REFERENCES


[1] Y. Yuan, C. Cronin, D. Müllensiefen, S. Fujii, and P. E. Savage, "Perceptual and automated estimates of infringement in 40 music copyright cases," 2023.

[2] K. Park, S. Baek, J. Jeon, and Y.-S. Jeong, "Music plagiarism detection based on siamese cnn," *Hum.-Cent. Comput. Inf. Sci*, vol. 12, pp. 12–38, 2022.

[3] T. He, W. Liu, C. Gong, J. Yan, and N. Zhang, "Music plagiarism detection via bipartite graph matching," *arXiv preprint arXiv:2107.09889*, 2021.

[4] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza, "A holistic approach to polyphonic music transcription with neural networks," *arXiv preprint arXiv:1910.12086*, 2019.

[5] C. Donahue and P. Liang, "Sheet sage: Lead sheets from music audio," *Proc. ISMIR Late-Breaking and Demo*, 2021.

[6] T. Deng, E. Nakamura, and K. Yoshii, "End-to-end lyrics transcription informed by pitch and onset estimation," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, ISMIR, 2022, pp. 633–639.

[7] I. Bukey, M. Feffer, and C. Donahue, "Just label the repeats for in-the-wild audio-to-score alignment," *arXiv preprint arXiv:2411.07428*, 2024.

[8] D. Malandrino, R. De Prisco, M. Ianulardo, and R. Zaccagnino, "An adaptive meta-heuristic for music plagiarism detection based on text similarity and clustering," *Data Mining and Knowledge Discovery*, vol. 36, no. 4, pp. 1301–1334, 2022.

[9] N. Borkar, S. Patre, R. S. Khalsa, R. Kawale, and P. Chakurkar, "Music plagiarism detection using audio fingerprinting and segment matching," in *2021 Smart Technologies, Communication and Robotics (STCR)*, IEEE, 2021, pp. 1–4.

[10] S. De, I. Roy, T. Prabhakar, *et al.*, "Plagiarism detection in polyphonic music using monaural signal separation," *arXiv preprint arXiv:1503.00022*, 2015.

[11] M. Marolt, "A mid-level melody-based representation for calculating audio similarity.," in *ISMIR*, Citeseer, 2006, pp. 280–285.

[12] J. Serra, X. Serra, and R. G. Andrzejak, "Cross recurrence quantification for cover song identification," *New Journal of Physics*, vol. 11, no. 9, p. 093 017, 2009.

[13] Z. Yu, X. Xu, X. Chen, and D. Yang, "Learning a representation for cover song identification using convolutional neural network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 541–545.

[14] 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)*, IEEE, 2022, pp. 616–620.

[15] X. Du, Z. Wang, X. Liang, H. Liang, B. Zhu, and Z. Ma, "Bytecover3: Accurate cover song identification on short queries," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[16] F. Liu, D. Tuo, Y. Xu, and X. Han, "Coverhunter: Cover song identification with refined attention and alignments," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2023, pp. 1080–1085.

[17] J. Urbano, J. Lloréns, J. Morato, and S. Sánchez-Cuadrado, "Melodic similarity through shape similarity," in *Exploring Music Contents: 7th International Symposium, CMMR 2010, Málaga, Spain, June 21-24, 2010. Revised Papers 7*, Springer, 2011, pp. 338–355.

[18] F. Karsdorp, P. van Kranenburg, and E. Manjavacas, "Learning similarity metrics for melody retrieval," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*, 2019, pp. 478–485.

[19] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[20] G. Weber, *Versuch einer geordneten Theorie der Tonsetzkunst*. B. Schott's Söhne, 1832, vol. 1.

[21] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[22] T. Kim and J. Nam, "All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2023, pp. 1–5.

[23] J. Zhao, G. Xia, and Y. Wang, "Beat transformer: Demixed beat and downbeat tracking with dilated self-attention," *arXiv preprint arXiv:2209.07140*, 2022.

[24] J.-Y. Wang and J.-S. R. Jang, "On the preparation and validation of a large-scale dataset of singing transcription," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 276–280.

[25] C. Donahue, J. Thickstun, and P. Liang, "Melody transcription via generative pre-training," in *ISMIR*, 2022.

[26] T.-P. Chen, L. Su, *et al.*, "Harmony transformer: Incorporating chord segmentation into harmony recognition," *Neural Netw*, vol. 12, p. 15, 2019.

[27] G. Koch, R. Zemel, R. Salakhutdinov, *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, Lille, vol. 2, 2015, pp. 1–30.

[28] Y. Li, R. Yuan, G. Zhang, *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2023.

[29] Y. Li, R. Yuan, G. Zhang, *et al.*, "Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning," *arXiv preprint arXiv:2212.02508*, 2022.