IADIS International Journal on WWW/Internet
Vol. 22, No. 2, pp. 75-92
ISSN: 1645-7641

LYRICCOVERS 2.0: AN ENHANCED DATASET FOR COVER SONG ANALYSIS

Maximilian Balluff[1], Maximilian Auch[1], Peter Mandl[1] and Christian Wolff[2]
[1]Hochschule München University of Applied Sciences, Germany
[2]University Regensburg, Germany

ABSTRACT

This research offers a detailed examination of a novel dataset that collates original musical compositions alongside their derivative cover versions. Unique in its inclusion of both links to YouTube as well as and lyrical content, the dataset enlists more than 78,000 tracks, encompassing more than 24,000 cover song groupings. It stands as the most diverse compendium of cover songs currently available for study. The characteristics of the *LyricCovers* dataset are thoroughly analyzed through its metadata, and empirical evaluations in the subsequent experimental lyrics analysis section suggest that lyrical analysis is a fundamental component in the identification and study of cover songs. This work presents a baseline approach to cover song detection, with an emphasis on lyrical content processing. It describes the extraction of lyrics from the audio files and the application of the Jina Embeddings 2 Model, fine-tuned with a hard triplet-loss objective, which successfully exploits lyric similarity to accurately identify cover songs.

KEYWORDS

Cover Song Detection, Music Information Retrieval, Dataset, Lyrics

1. INTRODUCTION

Cover versions, often referred to as cover songs, are artistic reinterpretations of previously recorded musical compositions by various artists (Magnus, 2022, p. 4). They occupy an important niche in the music industry and exert considerable influence on musical culture (Plasketes, 2016). These renditions are commonly encountered in diverse public spaces, including bars, clubs, festivals, and dining establishments. The proliferation of online music platforms has expanded the presence of cover songs, which frequently serve as background accompaniments or feature within the soundtracks of digital content (GEMA, 2021).

Accurate identification and subsequent monetization of cover songs present a formidable challenge for rights holders, leading to a growing demand for methods to reliably detect cover songs. Composers and performers, often affiliated with collecting societies, rely on precise song recognition mechanisms to equitably receive royalties (Serrà et al., 2010). Traditional automatic identification techniques perform adequately with original tracks; however, recognizing cover songs is significantly more complex due to the considerable variations that artists might infuse during their interpretive performances (Shazam, 2023)

In this paper, we contribute to the body of knowledge in two significant ways. First, we introduce *LyricCovers*, a comprehensive large-scale dataset of cover songs with annotated lyrics, substantially augmenting the available resources for music information retrieval research. Second, we propose a baseline approach for cover song detection that leverages the novel aspect of lyrical content analysis, harnessing advanced natural language processing techniques to enhance the accuracy of cover song identification.

2. RELATED WORK

The domain of Cover Song Detection (CSD) or Identification (CSI) has emerged as a pivotal component of Music Information Retrieval, engaging a growing number of researchers. The predominant focus of this research has been on the analysis of audio features. A seminal work by Tsai et al. (Tsai et al., 2005) established basic principles, employing tempo, key transposition, and accompaniment modifications to discriminate between cover songs and their originals. This direction was further developed by Serrà (2007) who introduced a cross-similarity matrix computation between the audio characteristics of the original tracks and their covers as a novel approach. Subsequent advances include the work of Du et al. (2021) who utilized convolutional neural networks (CNN) to create complex song embeddings that allow enhanced similarity assessments.

In contrast, the exploration of song lyrics within the realm of CSD has been considerably less extensive. An innovative study by Correya et al. (2018) leveraged the lyrics and title of songs using a bag-of-words model to detect relationships. Similarly, Vaglio et al. embarked on an integrative approach that merges lyrics analysis with audio features to identify cover songs (Vaglio et al., 2021). Their methodology bifurcates the audio processing into two distinct streams: lyrical and instrumental. The segment containing vocals is transcribed and analyzed using string-matching algorithms for cover identification. The synthesis of insights from both lyrical and musical analyses yields a comprehensive framework for determining cover song correlations.

The literature cites multiple datasets relevant to cover song research. A recent study (Balluff et al., 2024), highlights the *SecondHandSongs* dataset as one of the most widely used in the public domain (Bertin-Mahieux et al., 2011). This dataset comprises 18,196 songs organized into 5,854 cover clusters. Furthermore, the authors released the *Musixmatch* dataset, which contains a bag-of-words representation of song lyrics (Bertin-Mahieux et al., 2011; *musiXmatch Dataset, the Official Lyrics Collection for the Million Song Dataset*, n.d.).

The *SHS100K* dataset is the most substantial, built upon the *Million Songs Dataset*, encompassing 116,352 tracks with 9,359 identified cover groups (NovaFrost, 2017/2024). Numerous other datasets tailored for cover song identification exist, such as *Covers80* (Ellis, 2007), *Covers1000* (Tralie, 2017) and *DaTacos* (Yesiler et al., 2019). Typically, these repositories include pre-computed audio attributes and metadata. Among them, raw audio availability is confined to the *SecondHandSongs* compilations, which offer access via YouTube URLs. A notable limitation across these datasets is the absence of annotated lyrics.

3. DATASET

3.1 Description

Building upon our previous work LyricsCovers (Balluff et al., 2024), this dataset introduces significant improvements in data collection and processing. The dataset consists of a collection of original musical compositions along with their corresponding cover versions. For each musical piece, the dataset includes YouTube URLs, facilitating direct access to the audio content. Our improved download infrastructure, relocated from Amsterdam to Iowa, ensures more robust and reliable audio retrieval compared to the previous version. Additional under-the-hood optimizations have further enhanced the download process stability. As of November 2024, all audio files are accessible, though it is acknowledged that there exists a possibility for future unavailability should tracks be removed from YouTube.

In addition to the audio URLs, hyperlinks to annotated lyrics hosted on *genius.com* (*Genius |Song Lyrics & Knowledge*, n.d.) are provided, enabling researchers to examine the lyrical content of each song. Direct inclusion of lyrics within this dataset is precluded by copyright restrictions. Instead, by linking to external sources like Genius, we ensure that the dataset complies with copyright regulations, as the content remains hosted on platforms that manage the necessary licensing agreements.

Furthermore, during the creation of this dataset, we strictly adhered to the guidelines outlined in the robots.txt files of the respective websites. By following these protocols, we ensure that our data collection process respects the terms of service and web crawling permissions of platforms like Genius and YouTube.

To the best of the authors' knowledge, this corpus represents the first large-scale cover song repository that incorporates annotated lyrics. The dataset encompasses additional metadata, including but not limited to the language and genre of each track.

We are providing the complete source code not only for downloading the audio files and lyrics but also for reproducing the experiments. The code, along with a README file explaining how to run the experiments, is available for quick access on GitHub. The full dataset, as well as the subset used for training, is also included.

We have collected data from October 28 to November 11, 2023, utilizing the *genius.com* website, which offers a comprehensive catalog of musical works, replete with metadata and relational information concerning different versions of songs. The cover songs dataset comprises a total of 78,862 tracks, partitioned into 54,301 cover songs and 24,561 originals, thereby constituting one of the largest cover song datasets to date.

A key reason for creating a new dataset instead of extending an existing one is our emphasis on high-quality, reliable lyrics. Trusted sources like *genius.com* provide lyrics that have undergone community vetting and are less prone to errors compared to those produced by automated extraction methods. This ensures more accurate data for analysis. Our methodology prioritizes lyrics sourced from *genius.com* to establish a robust reference database, supplemented by automatically extracted lyrics when necessary.

3.2 Purpose

The task of detecting cover songs remains a formidable endeavor within the field of Music Information Retrieval. Despite notable progress in recent years that has led to substantial improvements in detection accuracy, the issue of accurately identifying cover songs has not been entirely resolved. Consequently, researchers continue to actively seek innovative methodologies aimed at further refining accuracy metrics. This paper introduces a novel dataset that offers an unprecedented feature, previously unavailable in both scope and form, to the research community. We contend that this dataset holds potential applications beyond cover song detection, specifically to advance the accuracy of lyrics transcription and refine genre classification algorithms within various music research domains.

3.3 Data Exploration and Analysis

3.3.1 Number of Covers per Original

For each original track in the dataset, there is a minimum of one cover version. On average, an original track has 3.21 cover renditions, with a standard deviation of 3.78. The median number of covers per original is 2. Certain tracks exhibit higher cover frequency; for example, the popular Christmas carol "Silent Night, Holy Night" boasts 145 cover versions. The distribution of the number of covers per original track is shown in Figure 1.

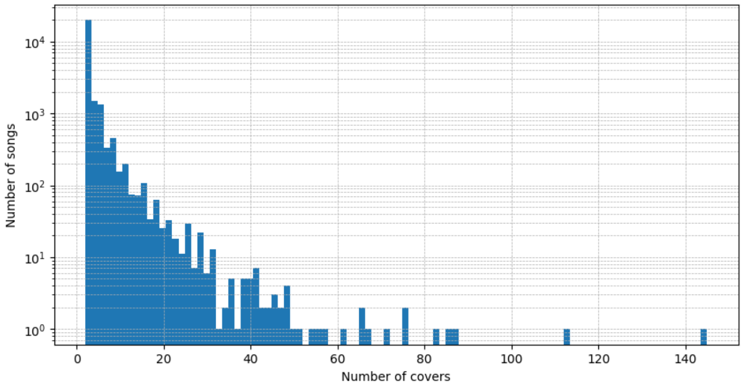

Figure 1. Cover songs per original songs

3.3.2 Languages

The dataset specifies the language for most of the songs, with only 859 (1.09%) tracks that do not specify the language. A total of 80 distinct languages are present within the collection. The five most common languages in the data set are English, containing 82.65%, followed by Spanish (3.46%), Portuguese (2.79%), Japanese (1.41%) and French (1.37%). Russian (1.23%), Italian (1.20%), German (1.19%), Korean (0.82%), and Hebrew (0.61%) round out the top ten languages. The language distribution for the cover songs mirrors the overall dataset, with minor variances. Figure 2 show the distribution of the top 10 languages per original and cover.

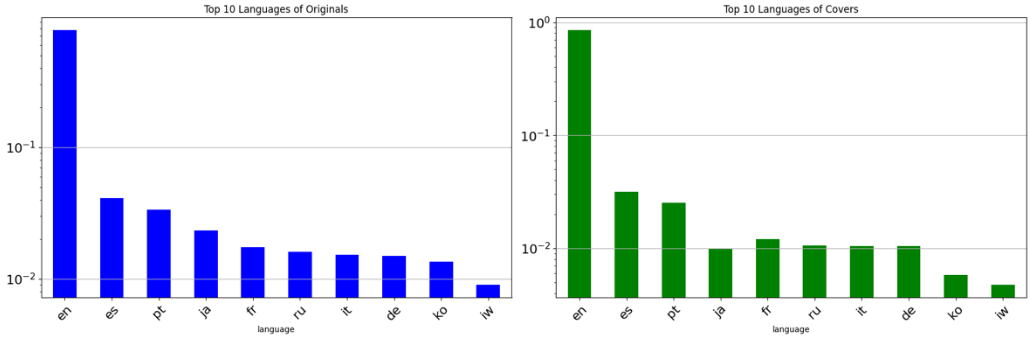

Figure 2. Top 10 languages per original and cover

About 95% of the cover songs are in the same language as their original, which is to be expected according to the definition of a cover. We analyzed the remaining 5% to see how the languages change. Unsurprisingly, English emerges as the primary target language for cross-language covers, with a significant number of songs originally in Japanese, Korean, Russian, French, German, Spanish, Portuguese, Hebrew, and Italian being covered in English. This pattern reflects the global dominance of the English-language music market and its role as a bridge for international music exchange.

Particularly noteworthy is the strong flow from Japanese and Korean originals to English covers, as shown in the Sankey diagram, suggesting a significant trend of East Asian music being adapted for English-speaking audiences. This corresponds to the rising global popularity of J-pop and K-pop over recent decades, where successful songs are frequently re-recorded in English to reach broader international audiences.

European languages show more diverse patterns of interchange. While English remains a common target language, we observe substantial flows between linguistically or geographically proximate languages - for instance, between German, French, and Italian. Spanish and Portuguese also display significant bidirectional flows, reflecting the close cultural ties within the Latin music sphere. These cross-language adaptations often involve more than simple translations, requiring careful consideration of rhythmic patterns, cultural references, and musical phrasing to maintain the song's artistic integrity while making it accessible to new audiences. The high percentage of cross-language covers (46%) indicates that language adaptation is a common practice in cover song creation, suggesting that artists frequently reinterpret songs across linguistic boundaries.

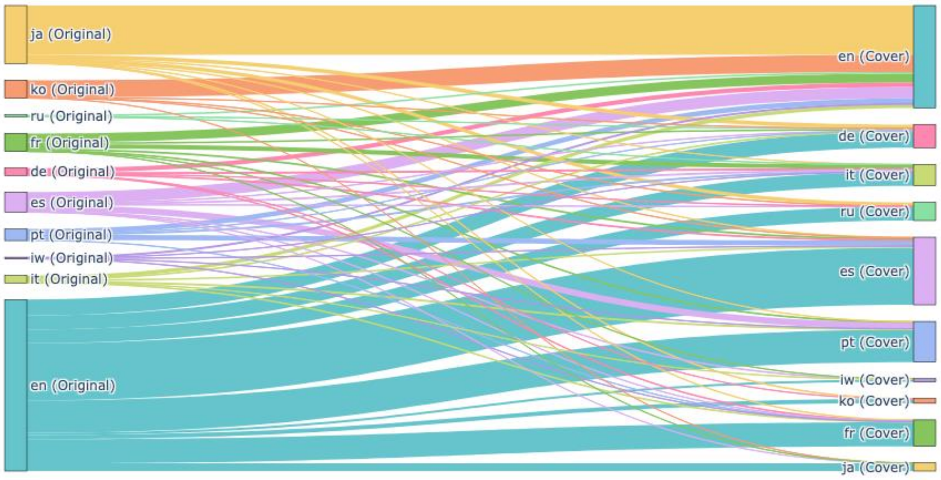

Figure 3. Language flow between original and cover songs

3.3.3 Artists

The dataset includes 9,325 unique artists responsible for the original songs, averaging 2.63 songs per artist and median of one and a 75 percentile of two songs per artist. For artists performing cover versions, the dataset comprises 18,963 distinct artists. The Beatles are the most covered artists in our Dataset with a total of 945 covers, they are followed by Bod Dylan (401) and Stevie Wonder (369), a detailed list of the top 10 original artist and cover artist can be found in Table 1. Top 10 original and cover artists.

Table 1. Top 10 original and cover artists

Top 10 Original Artists	Count	Top 10 Cover Artists	Count
The Beatles	154	KIDZ BOP Kids	941
Bob Dylan	117	Glee Cast	564
Taylor Swift	95	Vitamin String Quartet	537
Toby Fox	77	Scott Bradlee's Postmodern Jukebox	280
Frank Sinatra	72	The Grateful Dead	235
The Rolling Stones	68	First to Eleven	198
Metallica	67	NateWantsToBattle	170
Stevie Wonder	67	Cimorelli	166
David Bowie	61	Midnite String Quartet	164
Elvis Presley	58	J.Fla	140

Figure 4 illustrates the distribution of cover versions among the original artist. The majority of artists have one or two covers to their credit, whereas a small subset has accumulated more than 10 covers.

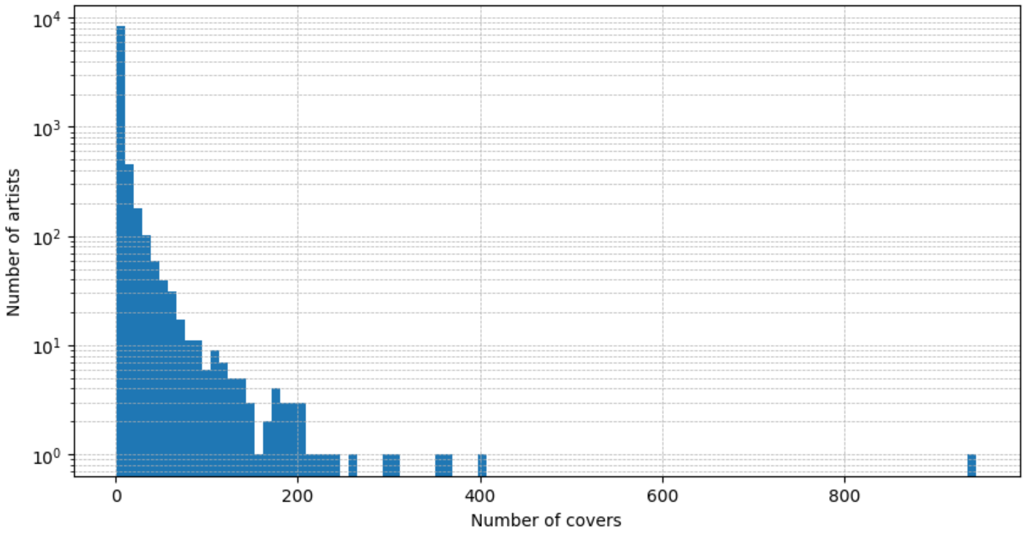

Figure 4. Covers per artist

3.3.4 Classification of Musical Genres

The task of delineating musical genres in the dataset derived from *genius.com* presents unique challenges. The platform does not assign a singular genre to each song; instead, it employs an extensive array of up to 24 tags. These tags are multifaceted, denoting not only the genre, but also aspects such as geographic origin, language, instrumental composition, and supplemental classifications such as "cover" or "live" performances. Additionally, there is a prevalence of composite tags that denote genre hybrids, such as "pop-rock".

After examining our dataset, which includes a total of 1,127 distinct tags, it was evident that the tags "pop," "cover" and "rock" were predominant. A more granular analysis of the top 50 tags shows that seven tags emerged as the most frequently occurring and genre-specific: "pop", "rock", "R&B", "country", "rap", "soul", and "jazz". A noteworthy disparity manifests between the prevalence of "pop" (47,882 instances) and "jazz" (3,034 instances), the latter occupying the seventh-most frequent position.

During the data preparation phase, we identified tags that could not be accurately classified within the predefined seven-genre framework. In instances where songs lacked a fitting tag after this categorization process, we assigned them to an "others" category. Our improved classification system now handles predefined genre mixtures more effectively, particularly common combinations like "pop-rock". This enhancement addresses a fundamental challenge in genre classification: music often blends different styles and rarely fits into a single category, much like mixed-media art. To account for this complexity, we established a "mixed genre" category for compositions that exhibit elements of multiple main genres. This systematic approach to handling hybrid genres has significantly improved our ability to accurately classify music that intentionally combines established styles.

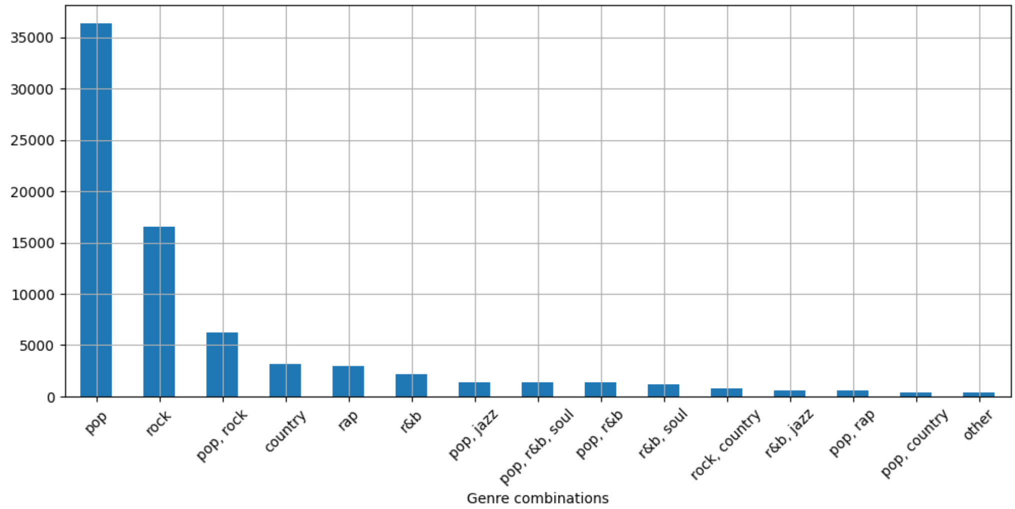

Figure 5. Top gerne (combination)

3.3.5 Temporal Distribution of Song Releases

The dataset encompasses a diverse range of songs, with original compositions spanning from the 18th century to the present day. Some of the timestamps are anomalous, such as the year "1" or the year "2024", a year beyond the dataset's scraping in 2023. Analyzing the temporal distribution of original songs, there is a relatively consistent presence of songs from the mid-1960s through to the early 2000s, accompanied by a marked increase in the 2010s, as depicted in Figure 6.

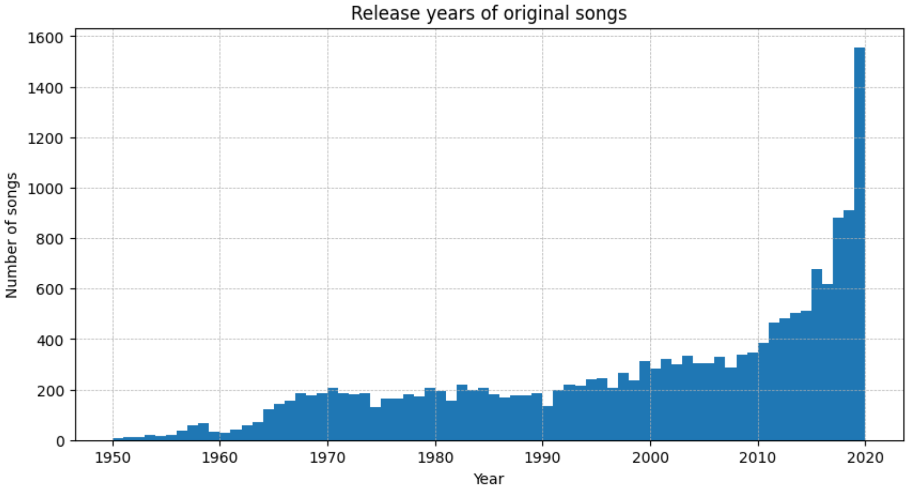

Figure 6. Temporal distribution of original song releases

In contrast, an examination of the temporal distribution pattern of cover songs reflects noticeable variability. During the 1950s, the phenomenon of covering was as prevalent as it is today. However, in contrast to the current era, many covers from that period were performed by lesser-known or obscure artists, resulting in a scarcity of recordings. Subsequently, fewer of these historic covers are available on contemporary platforms such as YouTube (Wilson, 2018). This dynamic shifted with the rise of hip-hop in the early 2000s. Due to these historical fluctuations, songs from earlier decades, with a particular under-representation of pre-1990s music, are less prevalent in our dataset (Wilson, 2018).

The data set shows a gradual increase in the production of title songs from the 1990s onwards, which peaked in the second half of the 2010s. Most of the covers included in our dataset originate from the period spanning from 2010 to the present day, a trend substantiated by the distribution plot depicted in Figure 7.

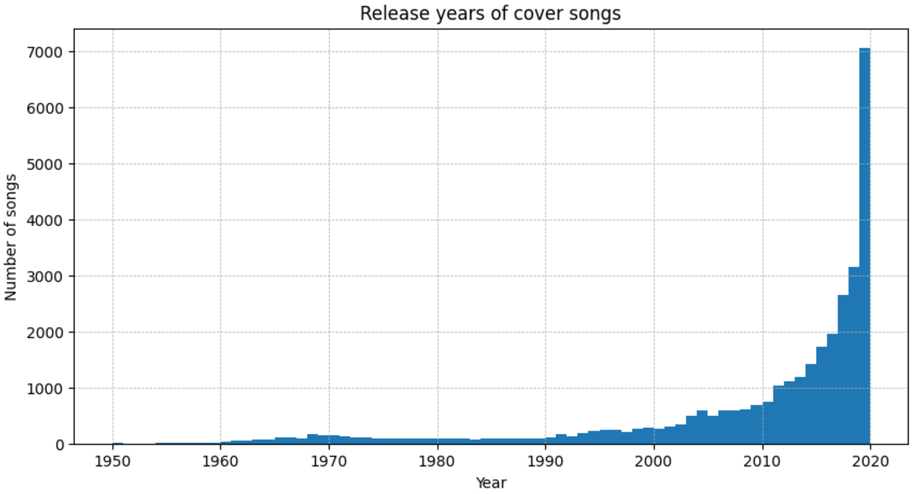

Figure 7. Temporal distribution of cover song releases

3.3.6 Enhanced Lyric similarity analysis

Building upon our previous research, we expanded our investigation into the lyrical relationships between cover songs and their original versions through a novel methodological approach. While our initial work employed conventional string similarity metrics such as Levenshtein and Jaro-Winkler (Prasetya et al., 2018), further analysis revealed the need for more sophisticated measurement techniques. Our key insight emerged from recognizing the parallel between cover songs and translations - both seek to preserve semantic content while potentially modifying structural elements. Drawing from this parallel, we incorporated metrics from machine translation evaluation into our analytical framework. The Word Error Rate (WER) proved particularly illuminating, as it captures the nuanced modifications - insertions, deletions, and substitutions - that characterize artistic reinterpretation in cover versions. Similarly, BLEU scores offered valuable insights into phrase-level preservation, revealing patterns in how cover artists maintain or modify sequential elements of the original lyrics.

To address aspects of similarity not captured by existing metrics, we developed the Vocabulary Overlap Score (VOS). This metric is expressed as

$$VOS = \frac{\left| V_{cover} \cap \mathrm{V}_{\mathrm{original}} \right|}{|V_{original}|}$$

where $\mathrm{V}_{\mathrm{cover}}$ and $\mathrm{V}_{\mathrm{original}}$ represent the respective sets of unique words in the cover and original versions. This normalized measure quantifies lexical preservation while accommodating creative adaptations, offering insights into how cover artists maintain vocabulary connections even when substantially reworking the original lyrics. The VOS complements traditional similarity metrics by specifically targeting vocabulary retention, an aspect previously underexplored in cover song analysis.

Our methodology encompasses a comprehensive evaluation of similarity metrics between original compositions and their cover versions, with several key methodological advances over previous approaches. The initial analysis establishes baseline similarity scores through direct comparison of original-cover pairs. However, we significantly extend this analysis by addressing the challenge of cross-language covers, which comprise about 10% of our dataset.

For cross-language comparisons, we developed a novel translation-based approach. From our dataset of approximately 50,000 cross-language pairs, we employed GPT-4o (gpt-4o-2024-08-06) for a random subset of 1000 pairs to translate cover versions into the language of their originals, enabling direct lyrical comparison. This method allows us to examine how lyrical content persists across linguistic boundaries while controlling for translation quality.

A major methodological advancement lies in our refined approach to control group construction. Our previous use of randomly selected unrelated songs proved insufficient for establishing meaningful baselines. Despite the inherent challenges of limited metadata, we developed two primary categories of control groups that provide more rigorous comparative frameworks.

The first category leverages artist-based relationships. We compare each cover song against the broader catalogue of the original artist, operating on the premise that artists typically maintain consistent stylistic elements, thematic preferences, and vocabulary choices across their work. This approach helps isolate the specific characteristics that distinguish cover versions from other songs within an artist's repertoire. Additionally, we examine relationships between different covers by the same covering artist, allowing us to identify artist-specific interpretation patterns and stylistic signatures in cover performances.

The second category employs temporal relationships, comparing songs based on their release years. This approach recognizes that contemporary songs often share stylistic elements, production techniques, and thematic concerns shaped by their cultural moment. Moreover, songs from the same period frequently exhibit similar structural characteristics influenced by prevailing genre conventions and technological capabilities of their era.

These methodologically robust control groups enable us to distinguish between similarity arising from cover relationship versus similarity stemming from shared temporal or artistic contexts. This framework provides a more nuanced understanding of how cover versions relate to their originals while controlling for confounding factors that might influence lyrical similarity measurements.

Table 2. Lyrics similarity metrics

	Bleu	WER	VOS	Levenshtein	Hamming	Jaro	Jaro-Winkler
Cover to Original	0.562	0.447	0.734	650.76	829.73	0.772	0.818
Cover to songs of original artist	0.034	1.270	0.267	1275.77	1620.53	0.689	0.670
Cover to songs of cover artist	0.043	1.217	0.359	1300.94	1685.81	0.737	0.736
Cover to songs of original production year	0.033	1.310	0.270	1420.06	1784.52	0.681	0.683
Cover to songs of cover production year	0.029	1.148	0.224	1437.70	1788.73	0.665	0.666
Translation of cover to Original	0.062	1.044	0.163	1362.34	1676.25	0.609	0.618

Table 2 presents comprehensive similarity metrics comparing cover songs to their originals and various control groups. For BLEU, VOS, Jaro, and Jaro-Winkler metrics, higher scores indicate stronger lyrical agreement, while for WER, Levenshtein, and Hamming distances, lower values signify better agreement. The direct comparison between covers and their originals shows notably strong relationships, with BLEU scores of 0.562, VOS of 0.734, and Jaro-Winkler similarity of 0.818. These scores consistently demonstrate that cover versions maintain substantial lyrical affinity with their original versions. All control group comparisons show markedly lower similarity scores, supporting our hypothesis that cover songs preserve significant textual elements from their originals.

Interestingly, when examining control groups, covers compared to other songs by the same cover artist show slightly higher similarity (VOS of 0.359) than comparisons to songs by the original artist (VOS of 0.267). This suggests that cover artists may maintain certain stylistic elements across their interpretations. Temporal comparisons, both using the original and cover production years, show lower similarities, indicating that era-specific patterns have less influence on lyrical content than artist-specific factors.

The analysis of translated covers presents unique insights. While these versions show moderate BLEU scores (0.062) and WER values (1.044), their overall similarity metrics are comparable to the control groups rather than to same-language covers. Our detailed examination reveals two key factors: First, some translations maintain semantic meaning while using substantially different vocabulary and phrasing. Second, in certain cases, automated translation services provided summary-like translations rather than direct conversions, likely due to copyright considerations (GEMA, 2024).

These findings suggest that while direct covers maintain strong lyrical fidelity to their originals, cross-language adaptations involve more substantial transformations that extend beyond simple translation.

4. EXPERIMENTAL COVER SONG DETECTION

As described in the previous section, lyrics can be a critical feature for the detection of cover songs. Consequently, this study introduces a basic approach using lyrics extracted from the songs. Owing to constraints in resources, the analysis was restricted to a subset of 44,593 tracks from the dataset.

Our core concept is to extract lyrics from cover song audio files and use them to find the most similar match within a reference database containing the original lyrics (transcribed and annotated). Therefore, we introduce the pipeline in Figure 8. First, the *Demucs* source separation algorithm (Défossez, 2022; Rouard et al., 2022) is applied to separate the vocal track from the audio files. The separation is followed by a transcription step using the *Whisper* tiny model to convert the vocal audio into text (Radford et al., 2022).The lyrics are then used to train an embedding model.

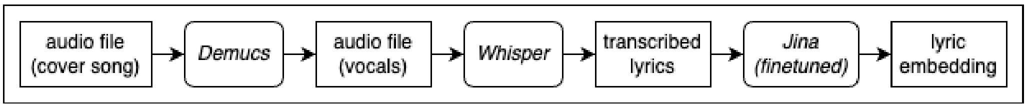

Figure 8. Audio processing pipeline

With the introduction of transformer models (Vaswani et al., 2023) this architecture has become the standard for many text processing tasks, such as classification, translation, or similarity. Our model is based on the *Jina Embeddings 2 Model* (jina-embeddings-v2-base-en) (Günther et al., 2024), a *AliBi*-based embedding model (Press et al., 2022) trained on 400 million sentence pairs to generate text embeddings. Unlike other BERT-based models, this model is able to handle long sequences of up to 8192 tokens, making it suitable for long texts (Günther et al., 2024). The model is fine-tuned for our downstream lyric similarity task using the online mining hard triplet loss function (Moindrot, 2018/2024; Musgrave et al., 2020; Schroff et al., 2015) with Euclidean distance. During training, the model is given batches of lyrics. For each batch, some of the lyrics belong to the same original song, such as the original song lyrics and the transcribed lyrics of the cover song, but it also contains unrelated lyrics. The goal is to learn embeddings with close distance for lyrics that belong to the same original while pushing away unrelated embeddings.

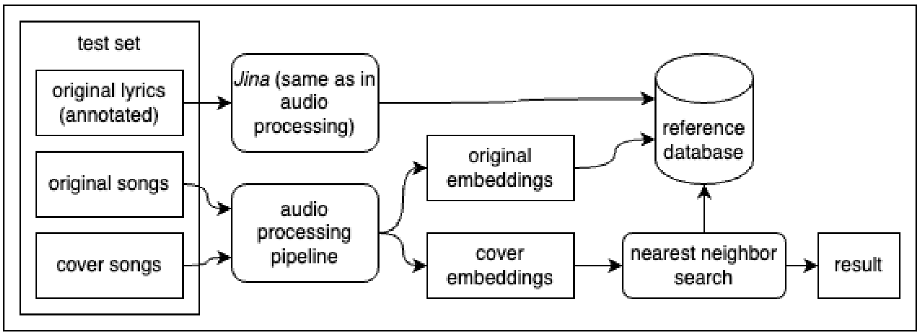

Figure 9. Inference/evaluation process

Our dataset is divided into three subsets: 60% for training, 20% for validation, and 20% for testing. Each subset contains a distinct set of original songs. Unlike a random split of the data, this method ensures that the model is not evaluated on the original songs it was trained on, thereby rigorously testing the generalization capabilities of the model embeddings. The dataset is distributed as follows: the training set includes 4,837 originals and 8,494 covers; the validation set comprises 1,612 originals and 2,417 covers; and the test set has 1,613 originals and 4,320 covers.

The training and validation subsets are used during model tuning. We employ learning rate reduction and early stopping techniques to adjust the learning rate during training and halt the process if the model performance does not improve, with the validation set loss being monitored.

Both annotated and transcribed lyrics of cover and original songs are used in training to increase the diversity and quantity of samples. The training process took approximately 5 hours, utilizing an Nvidia L4 GPU (24 GB memory) and a Google Cloud g2-standard-8 instance (8vCPU, 32 GB memory). Details of the training parameters are included in the source code.

Table 3. Evaluation results

Method	**mAP**	**MR**	**Precision@1**
Levenshtein	0.1033	2160	0.1491
Hamming	0.0107	2835	0.0149
Jaro	0.0008	4568	0.0137
Jaro-Winkler	0.0402	1856	0.0603
jina-embeddings-v2-small-en	**0.3963**	**631**	**0.5138**

Our approach represents a baseline method without optimization. Our evaluation uses the test set shown in Table 3. The reference database comprises annotated and transcribed lyrics from original songs, which are embedded using the fine-tuned *Jina* model. For the queries, only the transcribed lyrics of cover songs are used, which are similarly embedded by the *Jina* model and compared against the reference database via a nearest neighbor search. The mean average precision (mAP) and mean rank (MR) are commonly used for evaluation of cover song detection approaches (Chicco, 2021; Correya et al., 2018; Du et al., 2021, 2022). The mAP calculates the precision of the retrieved items, i.e. true matches in the first places of the ranking lead to a high mAP. The MR calculates the average rank of the first correct match, so lower values are better. Since this method has not yet been evaluated on other datasets, nor have other cover song detection methods been tested on this dataset, direct comparisons are not available at this stage. To provide a benchmark, we compared the results obtained from the *Jina* model against those derived using established string similarity metrics, mentioned previously. The results, as shown in Table 3, demonstrate that the fine-tuned *Jina* model outperforms the conventional similarity metrics, with the Levenshtein distance being the best among the string similarity measures.

4.1 Further Experiments

Our initial selection of models for the baseline approach was primarily driven by practical considerations such as open-source availability and operational simplicity. While this pragmatic approach enabled rapid development, we acknowledge that a more systematic model selection process could potentially yield improved cover song detection performance. To address this, we have begun a comprehensive evaluation of various model combinations for source separation and transcription. Our current experimental focus centers on quantifying the transcription performance of different pipeline configurations. We are evaluating the following models:

1. Source Separation Models:
   - Demucs with htdemucs configuration (Défossez, 2022; Rouard et al., 2022)
   - Spleeter (Hennequin et al., 2020)
2. Transcription Models:
   - Whisper tiny (Radford et al., 2022)
   - Whisper base (Radford et al., 2022)

These experiments are in their preliminary stages, and several model combinations remain to be evaluated. We plan to expand our investigation to include additional transcription models, such as Wav2Vec2 (Baevski et al., 2020), as well as more sophisticated versions of Whisper ("large" and "large_v3").

For our experimental protocol, we utilize a representative sample from our dataset, comprising [X] cover songs and [X] originals. For each song, we generate transcriptions using our processing pipeline with different model combinations. We then evaluate transcription quality by comparing the generated text against the reference lyrics using the BLEU score metric. This evaluation framework allows us to quantitatively assess the impact of different model selections on transcription accuracy. The preliminary results of these comparisons are presented in Table 4.

Table 4. Transcription Performance Comparison Across Model Combinations

Source separation model	Transcription model	BLEU score
Demucs	Whisper base	**0.168**
Demucs	Whisper tiny	0.125
Spleeter	Whisper tiny	0.092
No separation	Whisper tiny	0.119

The experimental results demonstrate that the combination of Demucs source separation with Whisper base achieves the highest BLEU score (0.168), significantly outperforming other configurations. Demucs consistently shows superior performance compared to Spleeter when paired with the same transcription model. Interestingly, direct transcription without source separation performs better than the Spleeter combination, suggesting that suboptimal source separation might be more detrimental than no separation at all.

The relatively low BLEU scores across all configurations (below 0.2) underscore the inherent challenges in lyric transcription from audio, particularly for cover songs where variations in musical style and arrangement can impact transcription accuracy. These findings inform our ongoing work to optimize the pipeline components and suggest that investing in more sophisticated models, particularly at the source separation stage, could yield meaningful improvements in transcription quality.

5. CONCLUSION

This study introduces *LyricCovers*, a large dataset of cover songs complemented by annotated lyrics, offering a novel avenue for exploration within the realm of Music Information Retrieval. Through meticulous data collection and detailed temporal analysis, we have compiled a resource of significant scale and diversity. This resource has the potential to improve research in various areas, including, but not limited to, cover song detection, lyric analysis, and genre identification. Our findings suggest that leveraging the LyricCovers dataset could enhance the performance of music recommendation systems by providing a more nuanced understanding of lyrical content and its variations in cover songs. Additionally, the dataset presents opportunities to refine existing music retrieval techniques, making them more robust in identifying covers and assessing their similarities to original tracks.

Beyond its intended purpose for cover song detection, *LyricCovers* has proven valuable for studying lyrical similarity, as our advanced experimental analysis shows. The employment of string similarity measures, translation metrics, and our newly developed Vocabulary Overlap Score (VOS) has indicated a markedly closer association between original songs and their covers compared to non-covers, with methodologically robust control groups based on artist relationships and temporal contexts providing more rigorous comparative frameworks. This comprehensive approach, which examines similarities against both artist catalogs and era-specific patterns while quantifying lexical preservation through VOS, underlines the dataset's capability to support investigations in the fidelity of lyrics for cover version detection.

This is also supported by our experimental cover song detection approach, which emphasizes lyrics as a discriminating feature. By integrating *Demucs* source separation algorithm and the robust transcription capabilities of the *Whisper* model, we effectively transcribed lyrics from audio tracks. Furthermore, the fine-tuning of the Jina Embedding Model with a triplet-loss objective demonstrated its potential to detect subtle similarities, suggesting that such methodological advancements could be applied to not only academic research but also industry applications, such as automated copyright monitoring and royalty distribution systems.

To conclude, *LyricCovers* stands as a comprehensive and innovative dataset that extends the frontier of cover song research, providing critical insights into the role of lyrics in music processing. The practical implications of this work underscore its potential to advance both academic research and practical applications within the music industry. As new methodologies are developed and tested using this dataset, it is expected that *LyricCovers* will significantly contribute to a deeper understanding of musical variations and their socio-cultural impacts.

6. FUTURE WORK

The evaluation of our baseline model underscores its performance in cover song detection and highlights the central role of lyrics in this area. However, this analysis also indicates potential areas for further improvement.

Currently, language discrepancies between the cover versions and the original versions present a significant challenge in our detection approach. While we initially considered adding language detection and translation steps to address this issue, our advanced similarity analysis of cross-language covers suggests that this might not be effective. The low similarity scores observed in translated lyrics, with BLEU scores of only 0.062 compared to 0.562 for same-language covers, indicate that cross-language adaptations involve substantial transformations beyond simple translation. This finding suggests that we need to explore alternative approaches for handling multi-language cover detection that can better capture these creative adaptations.

Furthermore, the quality of transcriptions can be inconsistent, influenced by factors such as the genre and language of the song. It is possible that the integration of modern large language models into the processing chain can eliminate transcription inaccuracies.

Future explorations can significantly enhance our findings. Firstly, combining lyrics with audio features, as demonstrated by Vaglio et al. (Vaglio et al., 2021), could improve model accuracy and robustness. Additionally, experiments on other datasets such as *SecondHandSongs* (NovaFrost, 2017/2024) will help validate the generalizability of our approaches and uncover dataset-specific trends.

Our current method proves to be computationally expensive due to the extensive processing required by the deep learning models involved in source separation, transcription, and language modeling. Each step requires significant processing time and requires costly hardware to perform efficiently. Addressing this issue is a high priority for future work. One potential solution is to eliminate the source separation step, as preliminary experiments with *Whisper* suggest that the transcription model achieves comparable results without this process. These findings are supported by our experimental results, where direct transcription without source separation achieved comparable BLEU scores (0.119) to configurations using Whisper tiny with source separation (0.125).

LyricCovers remains a promising resource for cover song identification research. It offers a robust foundation for future algorithmic advancements and will undoubtedly contribute to a deeper understanding of music consumption and artist influence. We are excited to see the innovative applications and discoveries that will emerge from the continued use and expansion of this dataset, and we welcome collaboration from the broader research community to explore its full potential.

ETHICAL STATEMENT

The *LyricCovers* dataset is intended to contribute positively to the music information retrieval task of cover song detection. It can aid in the advancement of detection to recognize the lineage of musical works. This may ensure that artists and right holders are accurately identified and compensated. We are aware that the dataset can be misused for copyright infringement when not handled responsibly. We strongly encourage users of the dataset to uphold copyright laws and respect intellectual property rights.

Data collection for this dataset adhered to the terms of service of platforms such as *genius.com* and *YouTube*. The dataset does not contain personal data, audio files, or lyrics, only references to sources, thus complying with copyright agreements. It is intended solely for research use and should not be commercialized.

USAGE OF GENERATIVE AI

In the writing process, we employed generative AI (Claude 3.5 Sonnet) strictly as an editorial aid to improve language and readability. All research, analyses, findings, and intellectual contributions are entirely our own work, with AI serving only to enhance the clarity of our presentation, similar to traditional editorial review.

REFERENCES

Balluff, M., Auch, M., Mandl, P. and Wolff, C., 2024. *A systematic mapping study for music cover detection* (manuscript submitted for publication).

Balluff, M., Mandl, P. and Wolff, C., 2024. Lyriccovers: A comprehensive large-scale dataset of cover songs with lyrics. *Proceedings of the International Conferences on Applied Computing and WWW/Internet 2024*.

Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B. and Lamere, P., 2011. The million song dataset. *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

Chicco, D., 2021. Siamese Neural Networks: An Overview. In H. Cartwright (ed.) *Artificial Neural Networks*. Springer US, pp. 73-94. https://doi.org/10.1007/978-1-0716-0826-5_3

Correya, A. A., Hennequin, R. and Arcos, M., 2018. *Large-Scale Cover Song Detection in Digital Music Libraries Using Metadata, Lyrics and Audio Features* (arXiv:1808.10351). arXiv. http://arxiv.org/abs/1808.10351

Défossez, A., 2022. *Hybrid Spectrogram and Waveform Source Separation* (arXiv:2111.03600). arXiv. https://doi.org/10.48550/arXiv.2111.03600

Du, X., Chen, K., Wang, Z., Zhu, B. and Ma, Z., 2022. Bytecover2: Towards Dimensionality Reduction of Latent Embedding for Efficient Cover Song Identification. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 616-620. https://doi.org/10.1109/ICASSP43922.2022.9747630

Du, X., Yu, Z., Zhu, B., Chen, X. and Ma, Z., 2021. Bytecover: Cover Song Identification Via Multi-Loss Training. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 551-555. https://doi.org/10.1109/ICASSP39728.2021.9414128

Ellis, D. P. W., 2007. *The covers80 cover song data set*. http://labrosa.ee.columbia.edu/projects/coversongs/covers80/

GEMA, 2021. *Diskotheken-Monitoring*. https://www.gema.de/musiknutzer/musik-lizenzieren/diskothekenmonitoring/

GEMA, 2024. *OpenAI: GEMA sues for fair compensation*. Gema.De. https://www.gema.de/en/news/ai-and-music/ai-lawsuit

Genius | Song Lyrics & Knowledge. (n.d.). Genius. Available at: https://genius.com/ (Accessed: 26 September 2024).

Günther, M. et al., 2024. *Jina Embeddings 2: 8192-Token General-Purpose Text Embeddings for Long Documents* (arXiv:2310.19923). arXiv. https://doi.org/10.48550/arXiv.2310.19923

Magnus, P. D., 2022. *A Philosophy of Cover Songs*. Open Book Publishers. https://doi.org/10.11647/obp.0293

Moindrot, O., 2024. *Omoindrot/tensorflow-triplet-loss* [Python]. https://github.com/omoindrot/tensorflow-triplet-loss (Original work published 2018)

Musgrave, K., Belongie, S. and Lim, S.-N., 2020. *PyTorch Metric Learning* (arXiv:2008.09164). arXiv. https://doi.org/10.48550/arXiv.2008.09164 *musiXmatch dataset, the official lyrics collection for the Million Song Dataset* (n.d.). Available at: http://millionsongdataset.com/musixmatch/ (Accessed: 26 September 2024)

NovaFrost, 2024. *NovaFrost/SHS100K* [Computer software]. https://github.com/NovaFrost/SHS100K (Original work published 2017)

Plasketes, G., 2016. *Play it Again: Cover Songs in Popular Music*. Routledge.

Press, O., Smith, N. A. and Lewis, M., 2022. *Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation* (arXiv:2108.12409). arXiv. https://doi.org/10.48550/arXiv.2108.12409

Radford, A. et al., 2022. *Robust Speech Recognition via Large-Scale Weak Supervision* (arXiv:2212.04356). arXiv. https://doi.org/10.48550/arXiv.2212.04356

Rouard, S., Massa, F. and Défossez, A., 2022. *Hybrid Transformers for Music Source Separation* (arXiv:2211.08553). arXiv. https://doi.org/10.48550/arXiv.2211.08553

Schroff, F., Kalenichenko, D. and Philbin, J., 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 815-823. https://doi.org/10.1109/CVPR.2015.7298682

Serrà, J., 2007. *Music similarity based on sequences of descriptors tonal features applied to audio cover song identification*, Master's Thesis, Barcelona, Spain, Universitat Pompeu Fabra.

Shazam, 2023. Shazam. https://www.shazam.com/de

Tralie, C. J., 2017. *Early MFCC And HPCP Fusion for Robust Cover Song Identification* (arXiv:1707.04680). arXiv. http://arxiv.org/abs/1707.04680

Tsai, W.-H., Yu, H.-M. and Wang, H., 2005. A Query-By-Example Technique for Retrieving Popular Songs with Similar Melodies. *ISMIR 2005, 6th International Conference on Music Information Retrieva*, 190.

Vaglio, A., Hennequin, R., Moussallam, M. and Richard, G. (2021). The words remain the same: Cover detection with lyrics transcription. *22nd International Society for Music Information Retrieval Conference ISMIR 2021*. https://hal.telecom-paris.fr/hal-03356164

Vaswani, A. et al., 2023. *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wilson, C., 2018. Is the Cover Making a Recovery? *Slate*. https://slate.com/culture/2018/10/cover-song-history-future-weezer-toto-africa.html

Yesiler, F. et al., 2019. *Da-Tacos: A dataset for cover song identification and understanding*. 8.