Information Research - Vol. 30 No. iConf (2025)

On the robustness of cover version identification models: a study using cover versions from YouTube

Simon Hachmeier and Robert Jäschke

DOI: https://doi.org/10.47989/ir30iConf47077

Abstract

Introduction. Recent advances in cover version identification have shown great success. However, models are usually tested on a fixed set of datasets which are relying on the online cover version database SecondHandSongs. It is unclear how well models perform on cover versions on online video platforms, which might exhibit alterations that are not expected.

Method. We annotate a subset of versions from YouTube sampled by a multi-modal uncertainty sampling approach and evaluate state-of-the-art cover version identification models.

Results. We find that existing models achieve significantly lower ranking performance on our dataset compared to a community dataset. We additionally measure the performance of different types of versions (e.g., instrumental versions) and find several types that are particularly hard to rank. Lastly, we provide a taxonomy of alterations in cover versions on the web.

Conclusions. We found that research in cover version identification shall be less dependent on SecondHandSongs but rather on more diverse datasets.

Introduction

In the context of western popular music, a cover version is a derivative of an original performance of a musical work. Artists perform versions to convey their subjective interpretations of musical works, which is a long-standing practice in musical culture. Usually, different versions of the same work share similar changes of individual notes (melody) or groups of notes (harmony) over time (Yesiler et al., 2021).

The research field of version identification (VI) deals with the automatic detection of cover versions in music collections. Recent approaches in VI aim to encode versions into representations retaining only relevant information in the context of cover versions (Du et al., 2021, 2022, 2023; Hu et al., 2022; Liu et al., 2023; Yesiler et al., 2020a, 2020b). For instance, Abrassart and Doras (Abrassart & Doras, 2022) report that melody, harmony, and lyrics are generally more relevant than rhythm. However, the actual relevance of each characteristic is non-trivial to predict and might strongly vary for different musical pieces. In contrast, characteristics irrelevant in the VI context are usually well agreed upon, such as the tempo or the key/scale.

Online video platforms feature various application scenarios for VI such as copyright infringement detection and music recommendation. Hence, the robustness of methods against noise and variance on the platform is important. One key peculiarity of VI in online videos is the alignment problem. In VI, this was addressed by summarization of musical content along the time axis including pooling mechanisms (Du et al., 2021, 2022; Yesiler et al., 2020a; Yu et al., 2020) and more recently by the matching of smaller chunks of the pairs (Du et al., 2023; Liu et al., 2023). Since YouTube is a collection of videos rather than versions (Except for YouTube's proprietary music streaming service *YouTube Music*), the relationship between videos and versions is an m-to-n relationship. This makes the alignment problem in online videos particularly challenging. For instance, a video might contain multiple versions (e.g., concert recordings) or only chunks of a version (e.g., guitar solo covers or tutorials (Hanson, 2018)). Additionally, videos might include noise such as commentary (e.g. people reacting to music (McDaniel, 2021)). Beside the alignment problem, other challenges might arise for VI in online videos such as the absence of the main melody (e.g. karaoke or instrumental versions (Agrawal & Sureka, 2013; Smith et al., 2017)), low fidelity in amateur recordings and versions occurring only in the background as accompaniment (Martet, 2016).

VI research has made great progress in recent years, mainly measured in metrics from MIREX (https://www.music-ir.org/mirex/wiki/2021:Audio_Cover_Song_Identification) and reported on community datasets like SHS100K-Test (Xu et al., 2018) and Da-Tacos (Yesiler et al., 2019). However, both of these datasets are based on the platform SecondHandSongs (SHS) (https://secondhandsongs.com/) curated by a community of volunteers (https://secondhandsongs.com/page/About) which makes present cover version collections subject to the selection policies of the platform. For example, *web covers* are considered an individual category of versions characterized by being released non-commercially (https://secondhandsongs.com/page/Guidelines/Entities/WebCover). At the same time, they appear to be less relevant for collaborators, since the amount of web covers is usually much lower than for commercially released covers as can be seen for the example ``Enter Sandman'' by Metallica (https://secondhandsongs.com/work/6616). What is more, due to a technical limitation of the application interface of SHS, the created datasets do not actually contain web covers. This poses the question whether VI models trained and evaluated on data from SHS are considering all relevant characteristics of versions and motivates our first research question:

- **RQ1:** do cover version datasets based on the platform SecondHandSongs represent the distributions of cover versions and their characteristics on YouTube?

We assume that there exists a subset of versions with specific characteristics on YouTube which are relevant in the context of VI but not found on the platform SHS: out-of-distribution data. Consequently, recent VI models are neither trained nor evaluated on data with regard to these characteristics. We therefore propose our second research question:

- **RQ2:** which characteristics of versions drive the uncertainty of existing VI models?

In this paper, we aim to explore the success and the challenges of VI on out-of-distribution data. Rather than relying on the cover version collection SecondHandSongs, we leverage the richness of creativity of the YouTube community. Applying a multi-modal uncertainty sampling approach, we identify the most uncertain version candidates. Subsequently, we obtain human annotations by workers on the crowdsourcing platform Mechanical Turk (MTurk). Lastly, two music experts curate a subset of the dataset and provide annotations of uncertainties in the problem context, together with a taxonomy of these.

In summary, the main contributions are:

- we provide a benchmark dataset SHS-YT (https://github.com/progsi/SHS-YT) created with a multi-modal uncertainty sampling approach followed by human annotations. It includes labels on an ordinal scale to reflect the complexity of VI on online video platforms (e.g., videos without musical content and identical audio tracks).
- two experts curate the provided dataset to gather insights into uncertainties in the VI context of online video platforms. We also provide a taxonomy extending an existing one (Yesiler et al., 2021).
- our benchmarks show that even the current state-of-the-art model under-performs on our proposed dataset. Additionally, we identify challenging alterations such as the isolation of single instruments or the vocal track which would be better addressed in the field of query-by-humming. This uncovers potential boundaries of cover version definitions.

Related work

Version identification datasets

VI datasets are composed of versions which are grouped by musical works. During training, VI models are optimized to encode audio representations of versions of the same work as similar and versions of different works as dissimilar. In the evaluation scenario, each version represents a query at a time and the remaining versions are ranked based on the musical similarity computed by the VI model. The resulting N rankings for a dataset with N versions serve as the input to compute retrieval metrics such as the mean average precision (MAP).

We provide an overview of the most popular datasets in VI which are used for benchmarking compared to the datasets used in this paper in Table 1. Recent VI approaches (Du et al., 2021, 2022, 2023; Hu et al., 2022; Liu et al., 2023; Yesiler et al., 2020a, 2020b) achieve MAP scores up to 0.96 on YouTubeCovers (Silva et al., 2015) and Covers80 (Ellis, 2011). The results on the larger datasets SHS100K-Test (Xu et al., 2018) and the Da-Tacos benchmark subset (Yesiler et al., 2019) are lower; for instance, CoverHunter (Liu et al., 2023) achieves the highest MAP but still does not surpass 0.90.

Dataset	Works	Versions
Covers80	80	160
Da-Tacos	1,000	15,000
Discogs-VI-YT-Test	9,878	116,547
SHS100K-Test	1,692	10,547
YouTubeCovers	50	350
SHS-SEED	100	2,404
SHS-YT	100	900
SHS-YT+2Q	100	1,092
SHS-YT+AllQ	100	3,289

Table 1. Popular VI benchmark datasets and the seed dataset and our annotated datasets in bold text

A commonality of all of these datasets but Covers80, is their utilization of SHS as a data source. The same accounts for the respective training sets of VI models: SHS100K-Train and the training subset of Da-Tacos (Yesiler et al., 2020a) which were used to train the recent VI models. Consequently, versions in the dataset can be found on YouTube, but are only included if these are manually collected by the SHS community. The question remains whether the distribution of variance of versions on YouTube is appropriately represented in existing benchmark datasets. A newer dataset, namely Discogs-VI-YT, is based on Discogs (https://www.discogs.com/) rather than SHS. It is the currently biggest dataset in VI. Since it is rather new, there are currently no benchmarks of the state-of-the-art VI models.

Music on YouTube

Various studies address the richness and diversity of versions on YouTube and corresponding classes. In Table 2 we distinguish between 4 classes of versions and provide some examples found in existing research. Liikkanen and Salovaara (2015) state that music is the most popular content type on YouTube. The results were derived from data about YouTube search trends, the most popular videos, and channels. The authors established twelve subclasses of versions segmented into three main classes: official (uploaded by copyright owners), user-appropriated (uploaded by fans) and derivative (e.g., cover versions). While the first two classes are expected to contain highly similar audio, the third class rather relies on music fans and hobby musicians. It includes stronger changes in musical characteristics; for instance, covers on instruments, parodies, or remixes.

The category of user-appropriated versions is also discussed by Martet (2016). The author also includes a new perspective on versions, including videos which contain versions rather as an accompaniment (e.g., for movie trailers).

Dataset	Works
Official	Official Music Video Professional Live Video
User-Appropriated	Lyric Video Slideshow
Cover	Guitar Cover Paraody Karaoke Version
Other	Tutorial Reaction Video

Table 2. Classes and examples of versions on YouTube.

From an application-driven perspective, studies have implemented pipelines to cope with copyright infringement detection (Agrawal & Sureka, 2013) and music retrieval (B. Li & Kumar, 2019; Smith et al., 2017) on YouTube. Smith et al. (2017) propose an approach processing audio, text, and video features to predict a version class. Similar to Airoldi et al. (2016) as well as Liikkanen and Salovaara (2015), the authors established classes like remixes and tutorials beside official music videos and live performances. Another approach to model classes of versions on YouTube derived clusters of categories of versions by a network analysis (Airoldi et al., 2016). The results emerged clusters corresponding to musical genres and situational contexts (Eg. covers and tutorials).

While the classes of versions in all of these studies might be relevant for VI research, their consideration in the field is rather limited. Yesiler et al. (2021) construct a taxonomy where they also mention some classes and the corresponding alterations of musical characteristics. To best of our knowledge, no existing benchmarks of VI models investigated the impact of the mentioned alterations on model robustness.

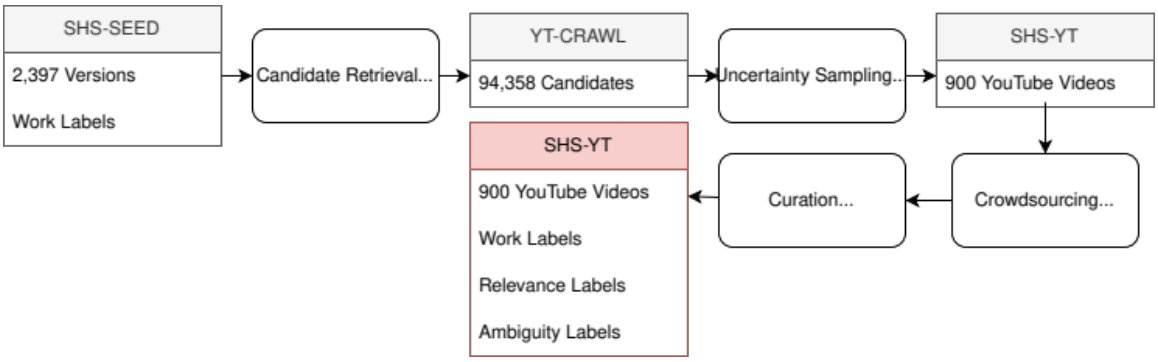

Figure 1. Dataset creation

Data creation

We here describe the steps of the creation process of the dataset SHS-YT as shown in Figure 1.

We aim to evaluate the performance of state-of-the-art VI models on out-of-distribution data. We select YouTube as a rich source for a diverse set of versions, since there are no constraints for uploaders as opposed to the policies on SHS.

To cover a representative subset of western popular music, we select the widely used SHS100K-Test as a seed dataset composed of works of western popular music. In particular, we choose the first 100 works from its test subset (https://github.com/NovaFrost/SHS100K/blob/master/SHS100K-TEST). These works are represented by 2,859 versions of which we successfully retrieved 2,397. We denote this dataset with 2,015 unique performers as *SHS-SEED*.

Candidate retrieval

The goal of the candidate retrieval step is to obtain a set of candidate versions to be included in our dataset. We apply an approach by Hachmeier et al. (2022) to formulate multiple text queries per work in SHS-SEED. We utilize the strings for performer and title of the first version of each work to formulate queries and additionally formulate new queries using YouTube search suggestions (The list of queries per work can be found in our repository). On average, we formulate 44 text queries per work resulting in 4,365 text queries. We retrieve metadata for the top 50 videos per query (https://pypi.org/project/youtube-search-python) and drop videos with a length of 10 minutes or more. We denote the resulting collection of 94,358 videos as YT-CRAWL. We download the audio files for all videos with a sampling rate of 22,050 Hz(https://github.com/yt-dlp) and extract CREMA features which we use in the next step (https://github.com/bmcfee/crema).

Uncertainty sampling

In the uncertainty sampling step, we aim to reduce the number of versions to a smaller subset because of two reasons: Firstly, we are limited in annotation capacity. Secondly, we aim to focus on out-of-distribution data and want to focus on versions with characteristics not common to be found on SHS. We leverage the modalities of audio (CREMA features) and text (YouTube metadata). For both domains, we use models based on deep learning as proxies. In theory, only the audio information is necessary to determine whether two versions are associated with each other. However, we use the text-based proxy to systematically find candidates where the VI proxy over- or underestimates the musical similarity.

Modality proxies

We use the pre-trained model Re-MOVE (Yesiler et al., 2020b) as a proxy in the audio/music domain which is one of the best approaches for VI at the time of dataset creation. The model processes CREMA features, which represent harmonic and melodic progressions, and encodes these into 256-dimensional embeddings. The Cosine similarity of a pair of embeddings represents their musical similarity. For the text domain, we use the entity matching model Ditto (Y. Li et al., 2020). The model is based on BERT (Devlin et al., 2018) and encodes pairs of textual entities into BERT embeddings and predicts a binary matching confidence. From the SHS100K-Train dataset we create a train, validation, and test set with a ratio of 3:1:1 as proposed by Y. Li et al. (2020) with each containing positive and negative pairs of YouTube videos in a 1:4 ratio. We gather the negative pairs by randomly sampling videos from another randomly selected work. We use all of the proposed data augmentation techniques and the best performing language model (RoBERTa) as described by Y. Li et al. (2020). We apply the best model checkpoint evaluated on the test set after 50~epochs for our matching task.

Since the inclusion of YouTube descriptions yielded inferior results (F1 score of 0.27 against 0.95) we solely process YouTube titles and channel names.

Similarity and matching confidence aggregation

For each candidate in YT-CRAWL we compute a similarity and matching confidence with the proxy models. Since each of the works is represented by multiple versions in SHS-SEED (24 on average) we must aggregate the pairwise similarities and model confidences. For a work i, a set of query versions from SHS-SEED Q_i and a candidate version c_{ij} from YT-CRAWL, we compute the musical similarity $S_m(c_{ij})$ as the arithmetic mean of the Cosine similarities of the Re-MOVE outputs of all

pairs (c_{ij}, q) for $q \in Q_i$. In a preliminary experiment on the validation dataset of SHS100K we validated the aggregation by the arithmetic mean as opposed to aggregation by maximum. We further compute the textual similarity $S_t(c_{ij})$ for the same pairs as the maximum matching confidence based on Ditto. We motivate this because candidates with non-matching metadata among the queries shall not have impact on the matching decision as long as at least one candidate in Q_i matches. This is especially of relevance in cases with translated version titles. For instance, the version title 'Tiempo de Verano' (Spanish for 'Time of the Summer' or 'Summertime') which is potentially a substring within a YouTube title might match the version title 'Summertime' with rather low confidence. Based on the aggregated values $S_m(c_{ij})$ and $S_t(c_{ij})$ for all candidates in YT-CRAWL, we conduct uncertainty sampling with two approaches: disagreement sampling and mutual uncertainty.

Disagreement sampling

We establish two disagreement groups: *DisagrAudio* denotes the candidates where the musical similarity high in contrast to the textual similarity and *DisagrText* represents the contrary case. We measure the disagreement as the absolute difference as shown in Table 3 and select the three candidates with the highest disagreement for both disagreement groups per work.

Group	Ranking Function
DisagrAudio	$S_m(c_{ij}) - S_t(c_{ij}) if S_m(c_{ij}) > S_t(c_{ij})$
DisagrText	$S_t(c_{ij}) - S_m(c_{ij}) if S_t(c_{ij}) > S_m(c_{ij})$
DisagrUnc	$-\left\|\left(S(c_{ij}), S^*(C_i)\right)\right\|_2$

Table 3. Uncertainty groups and their constraints. We sample the top three results returned by each ranking function.

Mutual uncertainty

We denote the mutual uncertainty group by *DisagrUnc* containing the top three candidates with the highest mutual uncertainty. Works with less than three candidates for DisagrAudio are filled with samples from this group as well. As shown in Table 3, we compute the mutual uncertainty as the negative Euclidean distance between the two-dimensional vector $S(c_{ij}) = \left[S_m(c_{ij}), S_t(c_{ij})\right]^T$ and the vector $S^*(C_i)$, representing the center of uncertainty based on all candidates for the work C_i, defined as follows:

$$(1) \quad S^*(C_i) = \begin{matrix} S_m^*(C_i) \\ S_t^*(C_i) \end{matrix}$$

with

$$(2) \quad S_\theta^*(C_i) = \frac{1}{2}\left(S_\theta^{min}(C_i) + S_\theta^{max}(C_i)\right)$$

where $\theta \in \{m, t\}$ and $S_\theta^{min}(C_i)$ and $S_\theta^{max}(C_i)$ return the minimum and maximum of the Cosine similarities or matching confidences for all the candidates in C_i, respectively. In the following, we describe our annotation process of the resulting nine candidates per work.

Annotation

We impose an ordinal scale of classes and obtain annotations by workers from Amazon's Mechanical Turk (MTurk) and inhouse-experts.

Relevance classes

Prior VI datasets solely consider the membership of a version to a work (binary label). Hence, each item in the dataset is expected to contain music. Further, the versions in the dataset of the same work are expected to be different regarding different aspects such as tempo or timbre (Yesiler et al., 2021). Both is not guaranteed when dealing with our retrieved candidates from YouTube, since videos are not even guaranteed to contain music. We construct four classes on an ordinal scale with respect to the relevance to the query version:

- *NoMusic*: candidate version does not contain music and is not relevant.
- *NonVersion:* candidate version does not contain music and is not relevant.
- *Version*: candidate version is derived of the same work as the query versions and therefore relevant.
- *Match*: candidate version includes (parts of) the exact same audio as the original they are derived of (*user-appropriated* videos). The version is relevant.

We represent each work i by a query version which is a random version from SHS-SEED. The goal of the annotation step is to gather annotations about the relevance between i and the candidate in the set. We denote the resulting set of 900 annotated versions as SHS-YT.

Crowdsourcing

We publish one human intelligence tasks (HITs) on MTurk per work with instructions and examples as shown in Figure 2. Each contains the query version, the nine candidates and a quality check candidate with a known answer *Version* or *NonVersion* based on the works and versions in SHS-SEED. To simplify the task, we specifically instruct that excerpts are sufficient (e.g., a medley is a *Version* if it contains an excerpt of the query).

The interface and manual presented to the workers can be found in our published dataset.

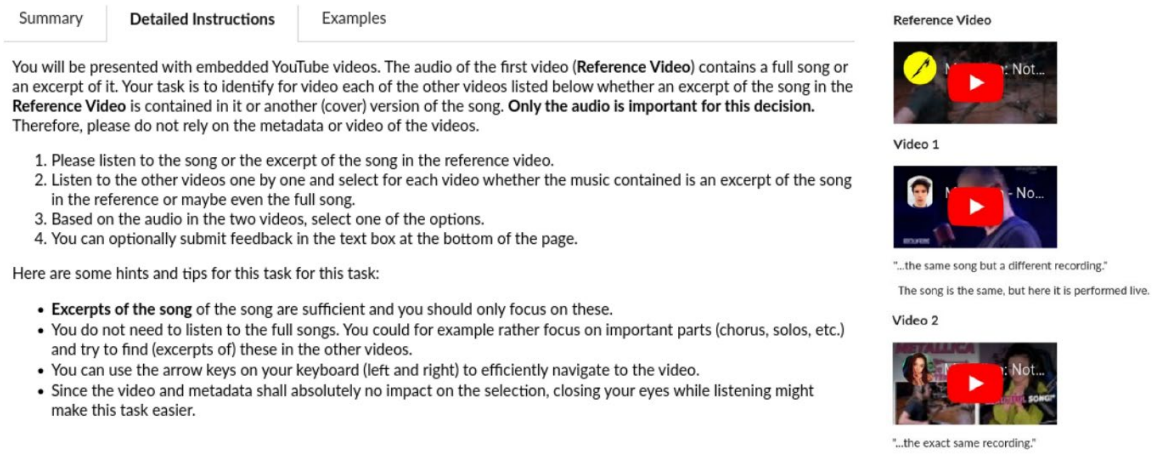

Figure 2. Our instructions and examples to workers as presented on MTurk. Please note that the examples of the right are cropped to fit

We measured the average time effort per annotation pair at 90~seconds and thus expect 15 minutes per HIT. We pay a reward of 3.2 US dollars per HIT corresponding to our domestic minimum wage, compensating our estimated time effort in consideration of the average currency exchange rate between our currency and the US dollar at annotation time.

We collect assignments by up to five workers per HIT. Following best practices to achieve annotation quality (Ghosh et al., 2019; Matherly, 2018; Mellis & Bickel, 2020; Peer et al., 2013) we only permit workers with more than 100 approved HITs and approval rate above or equal to 99%. We

reject assignments where workers fail the quality check or complete the assignments in less than ten seconds. In some cases, we accept assignments with failed quality checks due to proper justifications by workers. We do not include these assignments in our dataset. The final worker labels are obtained by majority voting: minimum three equal labels determine the decision for the label. Candidates which remain without a final label due to high variance in label responses are curated by the experts in the next step.

Curation

We employ two music experts for curation of the annotated dataset (The two experts have 15 years of musical experience on harmonic instruments.). The experts' task is to check the relevance labels of the workers for correctness, decide for a relevance label in undecided cases and to annotate the most prominent reason which makes a candidate more difficult to annotate (uncertainty class). In cases of uncertainty, the experts discuss decisions. Ultimately, experts and authors agreed upon including boundary cases as versions as well (e.g., remixes).

The first expert curates candidates labeled with *NoMusic* and 167 candidates with failed majority votes due to ties or shortage of worker assignments (because of failed assignment quality checks). Based on the collected reasons for uncertainties by the first expert, we formulate uncertainty classes and distinguish between uncertainties related to the version itself (e.g., *Song*: *Instrumental*) and uncertainties related to the Version in context of its occurrence in an online video (e.g., *Video*: *With Non-Music*). Some uncertainty classes just apply to one relationship class, for example, *Song*: *Same Artist* only applies if the candidate is a *NonVersion*. We provide a full documentation in our published repository.

The second expert utilizes the uncertainty classes directly and curates all candidates labeled with Version and the 96 most similar candidates labeled with *NonVersion* (Measured in mean Cosine similarity per benchmarking model as explained in the previous section) for error analysis. New uncertainty classes are collected and iteratively formulated, resulting in a total of 19 ambiguity classes. Based on these classes derived of observed examples, we construct a taxonomy of alterations.

Dataset analysis

Overview

We present the distributions of numerical YouTube attributes of SHS-YT in Figure 3. We observe a strong peak in duration around 3.5 minutes and in uploading dates between 2020 and 2022.

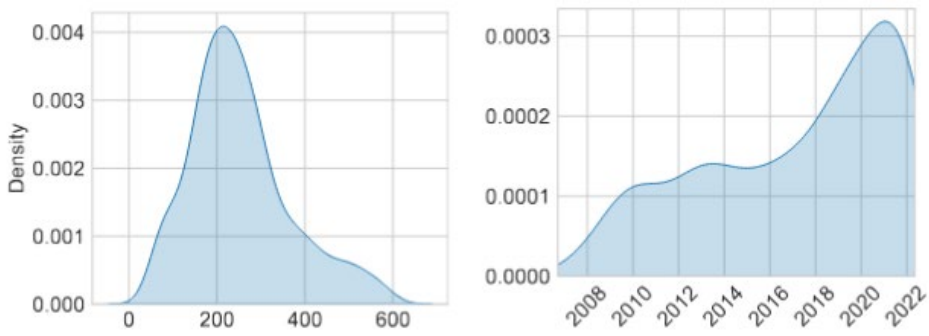

Figure 3. Gaussian kernel density estimates for duration (left) and uploading date (right) of the videos in the SHS-YT dataset. The bandwidth parameter is estimated by Scott's method

In Table 4 we show counts per annotation class and sampling group. The dataset mostly contains versions of other works than their respective query versions but also 197 versions of the same

works. *NoMusic* versions mostly occurred in the DisagrText group, which is expected, since a modeled musical similarity by Re-MOVE is rather unlikely with the absence of actual music. Similarly, the only 4 *Match* versions only occur in the DisagrAudio sampling group. SHS-YT contains 5 versions which are also contained in Da-Tacos; all are labeled with *NonVersion*. Regarding SHS100K, SHS-YT contains 5 versions from the test subset but from other works then in SHS-SEED, 2 from the validation subset and 13 candidates from the training subset. Hence, all of these candidates but one is annotated as *NonVersion*.

	Match	Version	NonVersion	NoMusic	Σ
DisagrAudio	4	89	200	0	293
DisagrText	0	82	142	76	300
DisagrUnc	0	26	280	1	307
Σ	4	197	622	77	900

Table 4. Uncertainty groups and their constraints. We sample the top three results returned by each ranking function

In Figure 4 we show the relative amounts of uncertainty classes excluding placeholder for 104 non-ambiguous versions according to the experts.

Non-musical content is the most represented uncertainty for *Versions* with 14% (n=77), followed by vocal-only. For *NonVersions*, the most frequent uncertainty is musical similarity between versions (*Song: Similar*) at 12%, followed by *NonVersions* from the same artist as the query version with 11%.

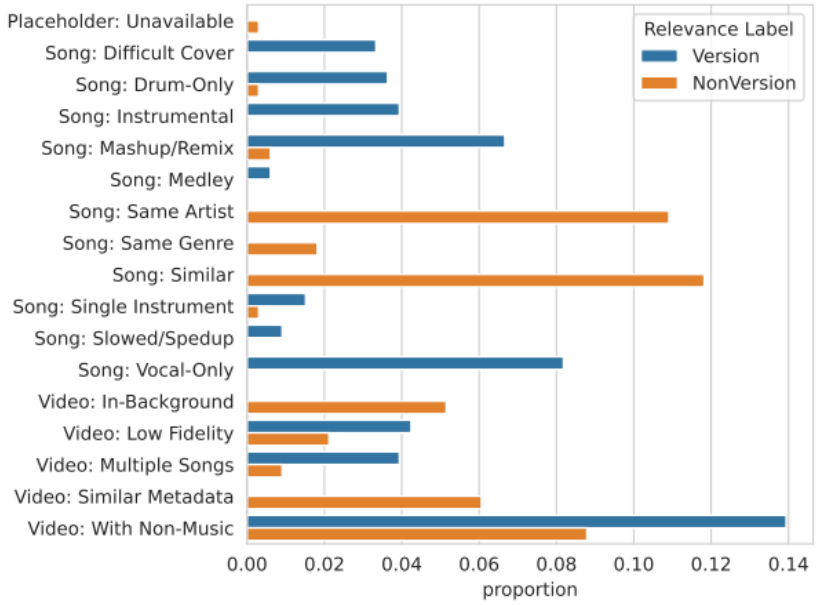

Figure 4. Relative proportion of uncertainty class annotated.

Annotation quality

Comparing the aggregated worker labels with expert labels for our 513 curated versions results in a Kendall's Tau of 0.81, indicating a strong positive association. However, the agreement among workers measured in Krippendorff's Alpha is just 0.43. The moderate level of inter-rater agreement might be partly due to the similarity of the VI task to the audio music similarity task which generally is associated with limited agreement as discussed in previous studies (Daikoku et al., 2020; Flexer et al., 2021; Flexer & Grill, 2016; Flexer & Lallai, 2019; Jones et al., 2007). Looking at the annotated uncertainty classes for candidates that are falsely labeled according to the expert (n=84) or which did not achieve a majority vote (n=167) uncovers some potential issues of workers. Especially versions which include non-musical and musical content seem to confuse workers (n=51). We found examples from 'The Voice' (https://youtu.be/aii62acsp_E) and a movie scene from 'Cocktail' (https://youtu.be/EFuBvEt84OI).

Benchmark

In this section, we conduct a benchmark on our proposed dataset with the goal to gather insights about the VI performance on out-of-distribution data. Since VI is a matching problem, we require relevant versions for all works in the dataset which is not the case for SHS-YT. Hence, we include versions from SHS-SEED. We construct two benchmark datasets derived of SHS-YT which we also show in Table 1. For both datasets, we exclude the versions which are included in the training and validation datasets of SHS100K:

- *SHS-YT+2Q*: SHS-YT with the query versions used for human annotation and one additional work from SHS-SEED. We select the version with the lowest version identifier which either is the original version or at least an earlier version derived of the original. This dataset includes minimum two relevant versions per work. In this dataset of 1,092 versions, our annotated versions account for around 82%. The 312 versions labeled as irrelevant (*NonVersion* and *NoMusic*) account for 29% of the dataset.
- *SHS-YT+AllQ*: Our proposed dataset with the all versions from SHS-SEED resulting in 3,289 versions. Here, our annotated versions account for around 27% of all versions. The 312 versions labeled as irrelevant account for 9% of the dataset.

Beside the modality proxies as described before, we evaluate two other VI models and a fuzzy matching baseline. CQTNet (Yu et al., 2020) is a VI model consisting of mainly convolutional neural networks. It processes constant-Q transform spectrograms (CQTs). The current state-of-the-art model is CoverHunter (Liu et al., 2023) which also processes CQTs but includes a conformer backbone (Gulati et al., 2020) and an attention mechanism (Okabe et al., 2018). The model is trained with a coarse-to-fine training scheme to address the alignment problem.

Both models are trained and validated on SHS100K. We use the pre-trained models provided by the authors. In contrast to the models we benchmark, the approach by Abrassart and Doras (Abrassart & Doras, 2022), LyraCNet (Hu et al., 2022) and the ByteCover models (Du et al., 2021, 2022, 2023) are not publicly available (https://github.com/Orfium/bytecover).

Overall performance

First, we evaluate the performance of models like in traditional VI research and consider only the binary label (relevant or not).

We report two evaluation metrics suggested by MIREX: (https://www.music-ir.org/mirex/wiki/2021:Audio_Cover_Song_Identification) mean average precision (MAP) and mean rank of the first relevant item (MR1). Since precision for the first 10 items is not a fair metric when having works with less than 10 relevant items, we omit this metric in our evaluation.

		SHS-YT+2Q		SHS-YT+AllQ		SHS100K-Test		Da-Tacos	
	Model	MAP	MR1	MAP	MR1	MAP	MR1	MAP	MR1
Audio	CoverHunter	0.52	44.5	0.83	8.1	0.86	11.9	0.85	12.2
	CQTNet	0.50	35.8	0.72	12.4	0.66	54.9	0.74	10.7
	Re-MOVE	0.40	86.9	0.56	18.5	0.53	38.0	0.52	38.0
Text	Ditto	0.39	73.78	0.78	18.5	-	-	-	-
	Fuzzy	0.24	101.3	0.46	14.3	-	-	-	-

Table 5. Benchmark results of VI models, the entity resolution model Ditto and Fuzzy which is the token set ratio from rapidfuzz (Bachmann, 2021)

In Table 5 we report the respective results on our benchmark datasets, SHS100K-Test and Da-Tacos. Please note that we exclude Discogs-VI-YT (Araz et al., 2024). since it was published after our experiments. Furthermore, it has to be noted that both of the evaluation metrics are sensitive to the dataset size which is not negligible (see Table 1). However, smaller dataset sizes usually promote a higher MAP and even though SHS-YT+2Q is a smaller dataset than the others, we observe a rather strong performance drop in MAP between -34% (CoverHunter) and -13% (Re-MOVE). The performance drop is less apparent for CoverHunter at SHS-YT+AllQ and the performance even increases compared to SHS100K-Test for the other VI models. While this is likely due to the larger number of versions from SHS-SEED we further look into the pairwise Cosine similarities for different pairwise relationships in the following section.

A closing remark on the overall evaluation is the potential influence of sampling bias to the performance of Re-MOVE and Ditto since these models are used as modality proxies in dataset creation.

Distributions of Cosine similarities

To support a more well-grounded verdict about the difference of distributions of versions in SHS-YT to versions on SHS and hence in datasets like SHS100K and Da-Tacos, we investigate the Cosine similarities of pairs of versions. A version from SHS-SEED can be considered a baseline version (*SHS-Version*). Our RQ1 aims to uncover whether existing VI models treat two SHS-Versions of the same work as more similar then an *SHS-Version* compared to a version from SHS-YT of the same work (*YT-Version*). Similarly, the question arises whether *NonVersions* from SHS-YT (YT-NonVersion) are more similar than other NonVersions from SHS-SEED (*SHS-NonVersion*): the former are versions in the same YouTube result sets (e.g. of the same artists) and the latter are random other versions.

In Table 6 we show statistics about the respective Cosine similarity distributions of SHS-Versions compared to other types of versions based on the relevance class. We observe that the similarities among SHS-Versions is significantly lower than their similarity to *YT-Versions*. Also, similarities of SHS-Versions of different works are significantly lower than their similarities to *YT-NonVersions*; but the corresponding effect size is lower. Both of these observations are likely a reason for less consistent rankings based on the tested VI models and hence the lower MAP scores observed in the previous section. Additionally, these insights substantiate an answer to RQ1 that in fact there exist different distributions of versions on SHS and YouTube.

Relevance Class	CoverHunter	CQTNet	Re-MOVE	Support
SHS-Version	0.88 ± 0.07	0.61 ± 0.13	0.62 ± 0.16	96,502
YT-Match	0.87 ± 0.08	0.61 ± 0.17	0.66 ± 0.19	44
YT-Version	**0.80 ± 0.10**	**0.48 ± 0.19**	**0.45 ± 0.24**	5,021
SHS-NonVersion	0.68 ± 0.04	0.33 ± 0.08	0.36 ± 0.09	5,637,128
YT-NonVersion	**0.72 ± 0.05**	**0.37 ± 0.09**	**0.41 ± 0.14**	14,305
YT-NoMusic	0.68 ± 0.07	0.23 ± 0.05	0.22 ± 0.05	1,810

Table 6. Arithmetic means and standard deviations of Cosine similarities between the *SHS-Versions* and the respective other versions. The prefix YT- indicates that the version is from SHS-YT and SHS- indicates that it is from SHS-SEED. Bold formatting indicates that means are statistically significant different measured with the Two-Sample-t-Test at $p < 0.01$

	Uncertainty Class	CoverHunter	CQTNet	Re-MOVE	Support
	SHS-Version	0.88 ± 0.07	0.61 ± 0.13	0.62 ± 0.16	96,502
YT-Version	*Version: Difficult Cover*	**0.82 ± 0.11**	**0.55 ± 0.17**	**0.55 ± 0.20**	293
	Version: Drum-Only	**0.72 ± 0.05**	**0.28 ± 0.07**	**0.23 ± 0.06**	321
	Version: Instrumental	**0.68 ± 0.12**	**0.38 ± 0.24**	**0.38 ± 0.28**	364
	Version: Mashup/Remix	**0.76 ± 0.07**	**0.44 ± 0.12**	**0.41 ± 0.21**	518
	Version: Medley	**0.72 ± 0.03**	**0.32 ± 0.09**	**0.25 ± 0.06**	86
	Version: Single Instrument	**0.80 ± 0.05**	0.68 ± 0.13	**0.46 ± 0.10**	195
	Version: Slowed/Sped-up	0.87 ± 0.05	**0.54 ± 0.14**	**0.43 ± 0.24**	63
	Version: Vocal-Only	**0.77 ± 0.04**	**0.38 ± 0.09**	**0.23 ± 0.07**	718
	Video: Low Fidelity	0.86 ± 0.09	**0.57 ± 0.16**	**0.49 ± 0.29**	292
	Video: Multiple Versions	**0.79 ± 0.09**	**0.49 ± 0.17**	**0.52 ± 0.21**	343
	Video: With Non-Music	**0.81 ± 0.10**	**0.48 ± 0.18**	**0.50 ± 0.23**	1,027
	SHS-NonVersion	0.68 ± 0.04	0.33 ± 0.08	0.36 ± 0.09	5,637,128
YT-NonVersion	*Version: Mashup/Remix*	**0.78 ± 0.04**	**0.42 ± 0.07**	0.33 ± 0.14	53
	Version: Same Artist	**0.76 ± 0.04**	**0.45 ± 0.08**	**0.51 ± 0.11**	862
	Version: Same Genre	**0.75 ± 0.05**	**0.40 ± 0.09**	**0.53 ± 0.13**	169
	Version: Similar Version	**0.76 ± 0.06**	**0.45 ± 0.08**	**0.51 ± 0.11**	1,069
	Video: Multiple Versions	**0.71 ± 0.05**	**0.39 ± 0.08**	**0.40 ± 0.14**	102

Table 7. Arithmetic mean and standard deviation of Cosine similarities between versions in SHS-SEED and a version from SHS-YT grouped by the uncertainty class. Bold formatting indicates that means are statistically significant different measured with the Two-Sample-t-Test at $p < 0.01$

To address RQ2, we investigate the differences of Cosine similarities for subsets of relevance classes grouped by their corresponding uncertainty classes in Table 7. Our imposed ordinal relevance classes also allow for analysis of similarities when dealing with highly similar versions (*YT-Match*) and versions without music (*YT-NoMusic*). Interestingly, the similarities are neither significantly higher nor lower than the similarities to other *SHS-Versions*. Regarding *NoMusic* versions, we can also see rather high similarities which indicates a lack of robustness of VI models. Almost all the

YT-Versions are significantly less similar compared to SHS-Versions ($p < 0.01$). The most challenging classes for all the models appear to be drum-only versions, instrumental versions, and medleys. While the latter is rather attributed to an alignment problem, the other two are most likely affected by the absence of the main melody and partly the harmony. Vocal-only versions which most likely only contain the main melody, appear to be hard for CQTNet and Re-MOVE and less so for CoverHunter. Difficulties for VI models for YT-NonVersions appear to arise due to versions being of the same artist, genre or just because they are similar by chance (*Version: Similar Version* and *Version: Mashup/Remix*).

In Figure 5, we further investigate the mean similarities by CoverHunter of different relevant versions. Apparently, the difficulty of drum-only versions is validated. We can also see that versions referring to multiple versions or including non-music noise impact the similarity. In the next section, we provide some examples for versions on YouTube which appear to be very challenging.

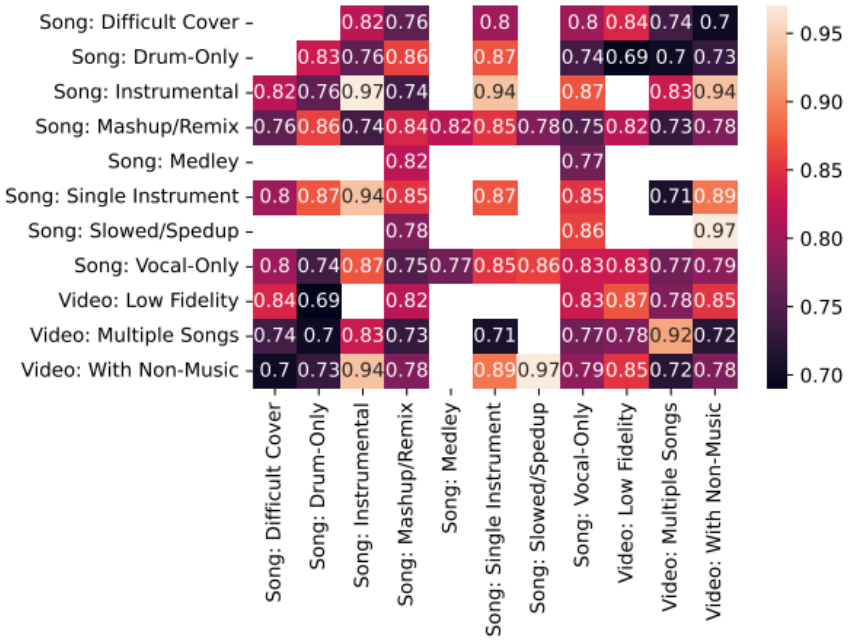

Figure 5. Mean Cosine similarities of CoverHunter embeddings between YT-Versions per respective uncertainty class

Error analysis

We examine the reasons for uncertainties more profoundly. First, we look at versions labelled *Non-Version* which are more similar than random other versions. We found that songs of the same genres are generally more similar, for instance bossa nova and blues. In theory VI models are not optimized to model genres per se. However, musical characteristics such as chord progressions (e.g., blues scheme) or rhythm (e.g., bossa nova beat) seem to be hard to disentangle from VI representations. Similarly, musical characteristics appear to correlate for versions of the same artists (e.g., Lady Gaga, Backstreet Boys, AC/DC). However, in some cases versions appear to be similar simply by similar chord progressions (e.g., 'Ultraviolence' (https://youtu.be/ZFWC4SiZBao) and 'Radioactive' (https://youtu.be/E5sVhFnrlTw)). Interestingly, we also found *NoMusic* versions with high similarity to Versions according to CoverHunter (https://youtu.be/PG6iJmbnOTY and https://youtu.be/svQD6mGDPXc). We assume that this is due to the matching of mute or low energy sections in these versions with mute parts of SHS-Versions after the alignment module.

Investigating some YT-Versions which appear to be difficult to detect, we found that vocal-only can refer either to versions with isolated voice stem by sound source separation (https://youtu.be/cixhJpyTWko) but also self-recorded vocal-only versions (https://youtu.be/24AKYyNusvs).

Discussion and implications

We summarize the findings gathered by our created dataset SHS-YT and the respective benchmarks. Regarding RQ1 we in fact confirmed a significant difference between some of the versions on YouTube and the ones included in SHS-based datasets. Based on our ordinal relevance labels, we derive that the difficulty especially arises due to relevant versions which are difficult to detect (false negatives) rather than irrelevant versions (false positives). However, some aspects such as similarity of songs within genres, of the same artists or with similar chord progressions seem to impact the overestimated similarity.

Looking at our dataset with annotated uncertainty classes reveals that the drum-only videos are rather challenging as well as instrumental versions. While the former do not include melody and harmony, these cases can be denoted as boundary cases. This raises the question about how a cover version is defined which is a question to be asked in musicology and maybe even of philosophical nature. Beside these rather song-specific uncertainty classes, there are also observable difficulties arising due to the alignment problem. While this is a general problem in VI, extreme cases such as medleys, multiple versions in a video and videos with versions and non-musical content still appear to be difficult for existing models.

To improve VI models in the future one solution is to rely on broader datasets in terms of data sources. For instance, by utilizing YouTube metadata to train weakly-supervised models. However, we propose another solution based on our observations. In Figure 6, we propose our taxonomy of cover versions in online videos. In the context of VI, musical characteristics which are discussed by Yesiler et al. (2021) (*Song* node) are one key component to model cover version relationships. Researchers are well aware about the importance of alterations in these characteristics and address them by augmentation techniques such as pitch-variations, tempo-variations. In the context of VI on YouTube (*Video* node), there are additional challenges which arise due to the context of online videos. Our observations provided examples for versions with low-fidelity and versions which occur in the background with foreground noise. We believe that both of these alterations can be well addressed by incorporating audio fingerprinting and noise mixing. We also found that isolated stems (e.g., drum-only, vocal-only versions) are particularly challenging. This is a problem which points to the related music information retrieval task of query-by-humming, where audio representations rely on single stems (usually the singing voice). In VI, an integration of sound source separation in augmentation techniques could further benefit model performance. Alternatively, rather than extracting the features in an end-to-end fashion using CQT spectrograms, one could extract features for melody, harmony, and rhythm separately like Abrassart & Doras (2022).

Lastly, the alignment problem which we have mentioned appears to be particularly strong on online video platforms. Not only can a version be represented only by a section (*Chunked*), but also along with multiple other versions or non-music noise.

The application of sliding time windows, possibly with different sizes followed by a maximum aggregation can address this problem. However, this might in turn increase the risk for false negatives and the computational load. We propose that also the synthesis of data by concatenation of different versions and non-musical noise such as commentary can help to make VI models more robust for these cases.

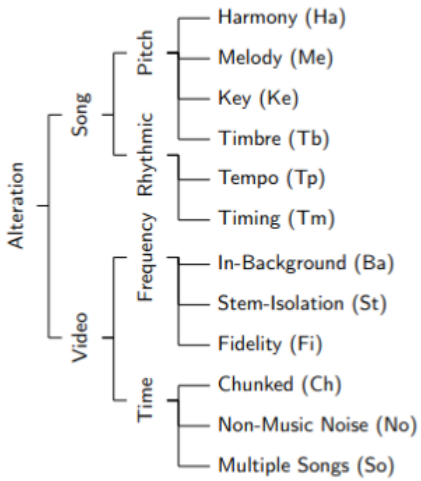

Figure 6. Taxonomy of cover versions in online videos

Conclusion and limitations

In this paper we proposed SHS-YT, a new benchmark dataset for VI. Created with a multi-modal uncertainty sampling approach and annotated by workers and experts including uncertainty classes, this dataset provided novel insights in the robustness of VI models. Lastly, we want to highlight some limitations of our study.

To the best of our knowledge, this is the first study which evaluates VI approaches with regard to different alterations among versions focusing on the most prominent uncertainty. Nevertheless, these classes might be partly subjective and cannot be fully isolated by other effects which might occur for certain pairs of versions. Due to the peculiarity of YouTube of being a dynamic online video platform, we cannot guarantee the presence of our videos on the platform in the future. In our repository, we provide all the URLs investigated. Due to copyright issues, we cannot provide the raw audio but only the extracted CQT and CREMA features. This paper focused on cover versions in the context western popular music. We are well aware that other genres might incorporate other characteristics which make this study less applicable. In future studies, the consideration of other genres with different characteristics could improve to gather an even broader overview of musical reinterpretations.

Acknowledgements

We thank the music experts for supporting the annotation processes.

About the authors

Simon Hachmeier is a Ph.D. student at the Berlin School of Library and Information Science at the Humboldt-Universität zu Berlin. He received his M.Sc. in Information Systems from the University of Innsbruck. Simon's research interest is cover version identification on the web. He can be contacted at simon.hachmeier@hu-berlin.de

Robert Jäschke is a Professor at the Berlin School of Library and Information Science at the Humboldt-Universität zu Berlin. He received his Ph.D. in Computer Science from the University of Kassel. Robert's research interests are web science and digital humanities. He can be contacted at robert.jaeschke@hu-berlin.de

References

Abrassart, M., & Doras, G. (2022). And what if two musical versions don't share melody, harmony, rhythm, or lyrics? International Society for Music Information Retrieval Conference.

Agrawal, S., & Sureka, A. (2013). Copyright Infringement Detection of Music Videos on YouTube by Mining Video and Uploader Meta-data. In V. Bhatnagar & S. Srinivasa (Eds.), Big Data Analytics (pp. 48–67). Springer International Publishing. https://doi.org/10.1007/978-3-319-03689-2_4

Airoldi, M., Beraldo, D., & Gandini, A. (2016). Follow the algorithm: An exploratory investigation of music on YouTube. Poetics, 57, 1–13. https://doi.org/10.1016/j.poetic.2016.05.001

Araz, R. O., Serra, X., & Bogdanov, D. (2024). Discogs-VI: A musical version identification dataset based on public editorial metadata. Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR).

Bachmann, M. (2021). Maxbachmann/RapidFuzz: Release 1.8. 0 (Vol. 10) [Computer software].

Daikoku, H., Ding, S., Sanne, U. S., Benetos, E., Wood, A. L., Fujii, S., & Savage, P. E. (2020). Human and automated judgements of musical similarity in a global sample. PsyArXiv Preprint.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint arXiv:1810.04805.

Du, X., Chen, K., Wang, Z., Zhu, B., & 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), 616–620. https://doi.org/10.1109/ICASSP43922.2022.9747630

Du, X., Wang, Z., Liang, X., Liang, H., Zhu, B., & Ma, Z. (2023). Bytecover3: Accurate cover song identification on short queries. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095389

Du, X., Yu, Z., Zhu, B., Chen, X., & Ma, Z. (2021). ByteCover: Cover Song Identification via Multi-Loss Training (arXiv:2010.14022). arXiv. https://doi.org/10.48550/arXiv.2010.14022

Ellis, D. P. W. (2011). The "covers80" cover song data set. http://labrosa.ee.columbia.edu/projects/coversongs/covers80

Flexer, A., & Grill, T. (2016). The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. Journal of New Music Research, 45(3), 239–251. https://doi.org/10.1080/09298215.2016.1200631

Flexer, A., & Lallai, T. (2019). Can we increase inter-and intra-rater agreement in modeling general music similarity? Conference of International Society for Music Information Retrieval (ISMIR), 494–500.

Flexer, A., Lallai, T., & Rašl, K. (2021). On evaluation of inter- and intra-rater agreement in music recommendation. Transactions of the International Society for Music Information Retrieval, 4(1), 182–194. https://doi.org/10.5334/tismir.107

Ghosh, S., Sperling, R., & Hooper, S. (2019). Using Amazon MTurk for research in academia: A beginner's guide for using Qualtrics, detecting VPN/proxy, limiting countries using geolocation & other tips. SSRN. https://doi.org/10.2139/ssrn.3455722

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & others. (2020). Conformer: Convolution-augmented transformer for speech recognition. arXiv Preprint arXiv:2005.08100.

Hachmeier, S., Jäschke, R., & Saadatdoorabi, H. (2022). Music Version Retrieval from YouTube: How to Formulate Effective Search Queries?. In *LWDA* (pp. 213-226).

Hanson, J. (2018). Assessing the educational value of YouTube videos for beginning instrumental music. Contributions to Music Education, 43, 137–158.

Hu, S., Zhang, B., Lu, J., Jiang, Y., Wang, W., Kong, L., Zhao, W., & Jiang, T. (2022). WideResNet with Joint Representation Learning and Data Augmentation for Cover Song Identification. Interspeech 2022, 4187–4191. https://doi.org/10.21437/Interspeech.2022-10600

Jones, M. C., Downie, J. S., & Ehmann, A. F. (2007). Human similarity judgments: Implications for the design of formal evaluations. International Society for Music Information Retrieval Conference (ISMIR).

Li, B., & Kumar, A. (2019). Query by video: Cross-modal music retrieval. International Society for Music Information Retrieval Conference (ISMIR).

Li, Y., Li, J., Suhara, Y., Doan, A., & Tan, W.-C. (2020). Deep entity matching with pre-trained language models. arXiv Preprint arXiv:2004.00584.

Liikkanen, L. A., & Salovaara, A. (2015). Music on YouTube: User engagement with traditional, user-appropriated, and derivative videos. Computers in Human Behavior, 50, 108–124. https://doi.org/10.1016/j.chb.2015.01.067

Liu, F., Tuo, D., Xu, Y., & Han, X. (2023). CoverHunter: Cover song identification with refined attention and alignments. 2023 IEEE International Conference on Multimedia and Expo (ICME), 1080–1085.

Martet, S. (2016). The circulation of user-appropriated music content on YouTube. YouTube and Music, 22(4), 169.

Matherly, T. (2018). A Panel for Lemons? Positivity bias, reputation systems and data quality on MTurk. European Journal of Marketing, 53. https://doi.org/10.1108/EJM-07-2017-0491

McDaniel, B. (2021). Popular music reaction videos: Reactivity, creator labor, and the performance of listening online. New Media & Society, 23(6), 1624–1641. https://doi.org/10.1177/1461444820918549

Mellis, A., & Bickel, W. (2020). Mechanical turk data collection in addiction research: Utility, concerns, and best practices. Addiction (Abingdon, England), 115. https://doi.org/10.1111/add.15032

Okabe, K., Koshinaka, T., & Shinoda, K. (2018). Attentive statistics pooling for deep speaker embedding. arXiv Preprint arXiv:1803.10963.

Peer, E., Vosgerau, J., & Acquisti, A. (2013). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. Behavior Research Methods, 46, 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

Silva, D. F., de Souza, V. M., & Batista, G. E. (2015). Music shapelets for fast cover song recognition. ISMIR, 441–447.

Smith, J. B. L., Hamasaki, M., & Goto, M. (2017). Classifying derivative works with search, text, audio, and video features. International Conference on Multimedia and Expo (ICME), 1422–1427. https://doi.org/10.1109/ICME.2017.8019444

Xu, X., Chen, X., & Yang, D. (2018). Key-invariant convolutional neural network toward efficient cover song identification. 2018 IEEE International Conference on Multimedia and Expo (ICME), 1–6. https://doi.org/10.1109/ICME.2018.8486531

Yesiler, F., Doras, G., Bittner, R. M., Tralie, C. J., & Serrà, J. (2021). Audio-based Musical Version Identification: Elements and Challenges. IEEE Signal Processing Magazine, 38(6), 115–136. https://doi.org/10.1109/MSP.2021.3105941

Yesiler, F., Serrà, J., & Gómez, E. (2020a). Accurate and scalable version identification using musically motivated embeddings. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 21–25. https://doi.org/10.1109/ICASSP40776.2020.9053793

Yesiler, F., Serrà, J., & Gómez, E. (2020b). Less is more: Faster and better music version identification with embedding distillation (arXiv:2010.03284). arXiv. https://doi.org/10.48550/arXiv.2010.03284

Yesiler, F., Tralie, C., Correya, A., Silva, D. F., Tovstogan, P., Gómez, E., & Serra, X. (2019). Da-TACOS: a dataset for cover song identification and understanding. Proc. of the 20th Int. Soc. for Music Information Retrieval Conf. (ISMIR), 327–334.

Yu, Z., Xu, X., Chen, X., & Yang, D. (2020). Learning a representation for cover song identification using convolutional neural network. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 541–545. https://doi.org/10.1109/ICASSP40776.2020.9053839