Audio Cover Song Identification and Similarity: Background, Approaches, Evaluation, and Beyond

Joan Serrà, Emilia Gómez, and Perfecto Herrera

1 Introduction

A cover version[1] is an alternative rendition of a previously recorded song. Given that a cover may differ from the original song in timbre, tempo, structure, key, arrangement, or language of the vocals, automatically identifying cover songs in a given music collection is a rather difficult task. The music information retrieval (MIR) community has paid much attention to this task in recent years and many approaches have been proposed. This chapter comprehensively summarizes the work done in cover song identification while encompassing the background related to this area of research. The most promising strategies are reviewed and qualitatively compared under a common framework, and their evaluation methodologies are critically assessed. A discussion on the remaining open issues and future lines of research closes the chapter.

1.1 Motivation

Cover song identification has been a very active area of study within the last few years in the MIR community, and its relevance can be seen from multiple points of view. From the perspective of audio content processing, cover song identification yields important information on how musical similarity can be measured and modeled. Music similarity is an ambiguous term and, apart from musical facets themselves, may also depend on different cultural (or contextual) and personal (or subjective) aspects [24]. The purpose of many studies is to define and evaluate the concept of music similarity, but there are many factors involved in this problem, and

Joan Serrà · Emilia Gómez · Perfecto Herrera
Music Technology Group, Department of Information and Communication Technologies, Universitat Pompeu Fabra. Tànger 122-140, office 55.318, 08018 Barcelona Spain
e-mail: `{joan.serraj,emilia.gomez,perfecto.herrera}@upf.edu`

[1] We use the term cover or version interchangeably.

Z.W. Raś and A.A. Wieczorkowska (Eds.): Adv. in Music Inform. Retrieval, SCI 274, pp. 307–332.
springerlink.com

some of them (maybe the most relevant ones) are difficult to measure [6]. Still, the relationship between cover songs is context-independent and can be qualitatively defined and objectively measured, as a "canonical" version exists and any other rendition of it can be compared to that.

The problem of identifying covers is also challenging from the point of view of music cognition, but apparently it has not attracted much attention by itself. When humans are detecting a cover, they have to derive some invariant representation of the whole song or maybe of some of its critical sections. We do not know precisely what is the essential information that has to be encoded in order for the problem to be solved by human listeners. Nevertheless, it seems relevant the knowledge gained about the sensitivity or insensitivity to certain melodic transformations, for example [17, 87]. In addition, when the cover is highly similar in terms of timbre, it seems that this cue can do the job to help us to identify the song even using very short snippets of it [71]. An additional issue that is called for by cover identification is that of the memory representation of the songs in humans. It could be either the case that the canonical song acts as a prototype for any possible version, and that the similarity of the covers is computed in their encoding step, or either that all the songs are stored in memory (as exemplary-based models would hypothesize) and their similarity is computed at the retrieval phase. For example, Levitin [46] presents evidence in favor of absolute and detailed coding of song specific information (at least for the original songs). On the other hand, Deliege [14] has discussed the possibility of encoding processes that abstract and group by similarity certain musical cues.

From a commercial perspective, it is clear that detecting cover songs has a direct implication to musical rights' management and licenses. Furthermore, quantifying music similarity is key to searching, retrieving, and organizing music collections. Nowadays, online digital music collections are in the order of ten [59] to a few hundred million tracks[2] and they are continuously increasing. Therefore, one can hypothesize that the ability to manage this huge amount of digital information in an efficient and reliable way will make the difference in tomorrow's music-related industry [10, 85]. Personal music collections, which by now can easily exceed the practical limits on the time to listen to them, might benefit as well from efficient and reliable search and retrieval engines.

From a user's perspective, finding all versions of a particular song can be valuable and fun. One can state an increasing interest for cover songs just by looking at the emergence of related websites, databases, and podcasts in the internet such as Second Hand Songs[3], Coverinfo[4], Coverville[5], Midomi[6], Fancovers[7], or YouTube[8].

[2] See for example http://www.easymp3downloader.com/, http://blog.wired.com/music/2007/04/lastfm_subscrip.html, or http://www.qsrmagazine.com/articles/news/story.phtml?id=5852.

[3] http://www.secondhandsongs.com

[4] http://www.coverinfo.de

[5] http://www.coverville.com

[6] http://www.midomi.com

[7] http://www.fancovers.com

[8] http://www.youtube.com

Frequently, these sites also allow users to share/present their own (sometimes home-made) cover songs, exchange opinions, discover new music, make friends, learn about music by comparing versions, etc. Thus, cover songs are becoming part of a worldwide social phenomena.

1.2 Types of Covers

Cover songs were originally part of a strategy to make profit from 'hits' that had achieved significant commercial success by releasing them in other commercial or geographical areas without remunerating the original artist or label. Little promotion and highly localized record distribution in the middle of the 20th century favored that. Nowadays, the term has nearly lost these purely economical connotations. Musicians can play covers as a homage or a tribute to the original performer, composer or band. Sometimes, new versions are rendered for translating songs to other languages, for adapting them to a particular country/region tastes, for contemporizing old songs, for introducing new artists, or just for the simple pleasure of playing a familiar song. In addition, cover songs represent the opportunity (for beginners and consolidated artists) to perform a radically different interpretation of a musical piece. Therefore, today, and perhaps not being the proper way to name it, a cover song can mean any new version, performance, rendition, or recording of a previously recorded track [42].

Many distinctions between covers can be made (see [27, 79, 89] for some MIR-based attempts). These usually aim at identifying different situations where a song was performed in the context of mainstream popular music. Considering the huge amount of tags and labels related to covers (some of them being just buzzwords for commercial purposes), and according to our current understanding of the term cover version, we advocate for a distinction based on musical features instead of using commercial, subjective, or situational tags. But, just in order to provide an overview, some exemplary labels associated with versions are listed below [42].

- Remaster: Creating a new master for an album or song generally implies some sort of sound enhancement (compression, equalization, different endings, fade-outs, etc.) to a previous, existing product.
- Instrumental: Sometimes, versions without any sung lyrics are released. These might include karaoke versions to sing or play with, cover songs for different record-buying public segments (e.g. classical versions of pop songs, children versions, etc.), or rare instrumental takes of a song in CD-box editions specially made for collectors.
- Live performance: A recorded track from live performances. This can correspond to a live recording of the original artist who previously released the song in a studio album, or to other performers.
- Acoustic: The piece is recorded with a different set of acoustical instruments in a more intimate situation.
- Demo: It is a way for musicians to approximate their ideas on tape or disc, and to provide an example of those ideas to record labels, producers, or other artists.

Musicians often use demos as quick sketches to share with band mates or arrangers. In other cases, a songwriter might make a demo in order to be send to artists in hopes of having the song professionally recorded, or a music publisher may need a simplified recording for publishing or copyright purposes.
- Duet: A successful piece can be often re-recorded or performed by extending the number of lead performers outside the original members of the band.
- Medley: Mostly in live recordings, and in the hope of catching listeners' attention, a band covers a set of songs without stopping between them and linking several themes.
- Remix: This word may be very ambiguous. From a 'traditionalist' perspective, a remix implies an alternate master of a song, adding or subtracting elements, or simply changing the equalization, dynamics, pitch, tempo, playing time, or almost any other aspect of the various musical components. But some remixes involve substantial changes to the arrangement of a recorded work and barely resemble the original one. Finally, a remix may also refer to a re-interpretation of a given work such as a hybridizing process simultaneously combining fragments of two or more works.
- Quotation: The incorporation of a relatively brief segment of existing music in another work, in a manner akin to quotation in speech or literature. Quotation usually means melodic quotation, although the whole musical texture may be incorporated. The borrowed material is presented exactly or nearly so, but is not part of the main substance of the work.

1.3 Involved Musical Facets

With nowadays' concept of cover song, one might consider the musical dimensions in which such a piece may vary from the original one. In classical music, different performances of the same piece may show subtle variations and differences, including different dynamics, tempo, timbre, articulation, etc. On the other hand, in popular music, the main purpose of recording a different version can be to explore a radically different interpretation of the original one. Therefore, important changes and different musical facets might be involved. It is in this scenario where cover song identification becomes a very challenging task. Some of the main characteristics that might change in a cover song are listed below:

- Timbre: Many variations changing the general color or texture of sounds might be included into this category. Two predominant groups are:
  - Production techniques: Different sound recording and processing techniques (e.g. equalization, microphones, dynamic compression, etc.) introduce texture variations in the final audio rendition.
  - Instrumentation: The fact that the new performers can be using different instruments, configurations, or recording procedures, can confer different timbres to the cover version.

- Tempo: Even in a live performance of a given song from its original artist, tempo might change, as it is not so common to control tempo in a concert. In fact, this might become detrimental for expressiveness and contextual feedback. Even in classical music, small tempo fluctuations are introduced for different renditions of the same piece. In general, tempo changes abound (sometimes on purpose) with different performers.
- Timing: In addition to tempo, the rhythmical structure of the piece might change depending on the performer's intention or feeling. Not only by means of changes in the drum section, but also including more subtle expressive deviations by means of swing, syncopation, pauses, etc.
- Structure: It is quite common to change the structure of the song. This modification can be as simple as skipping a short 'intro', repeating the chorus, introducing an instrumental section, or shortening one. But it can also imply a radical change in the musical section ordering.
- Key: The piece can be transposed to a different key or tonality. This is usually done to adapt the pitch range to a different singer or instrument, for 'aesthetic' reasons, or to induce some mood changes on the listener.
- Harmonization: While maintaining the key, the chord progression might change (adding or deleting chords, substituting them by relatives, modifying the chord types, adding tensions, etc.). This is very common in introduction and bridge passages. Also, in instrument solo parts, the lead instrument voice is practically always different from the original one.
- Lyrics and language: One purpose of performing a cover song is for translating it to other languages. This is commonly done by high-selling artists to be better known in large speaker communities.
- Noise: In this category we consider other audio manifestations that might be present in a song recording. Examples include audience manifestations such as claps, shouts, or whistles, audio compression and encoding artifacts, speech, etc.

Notice that, in some cases, the characteristics of the song might change, except, perhaps, a lick or a phrase that is on the background, and that it is the only thing that reminds of the original song (e.g. remixes or quotations). In these cases, it becomes a challenge to recognize the original song, even if the song is familiar to the listener. Music characteristics that may change within different types of covers are shown in table 1.

1.4 Scientific Background

In the literature, one can find plenty of approaches addressing song similarity and retrieval, both in the symbolic and the audio domains[9]. Within these, research

[9] As symbolic domain we refer to the approach to music content processing that uses, as starting raw data, symbolic representations of musical content (e.g. MIDI or **kern files, data extracted from printed scores). Contrastingly, the audio domain processes the raw audio signal (e.g. WAV or MP3 files, real-time recorded data).

Table 1 Musical changes that can be observed in different cover song categories. Stars indicate that the change is possible, but not necessary.

	Timbre	Tempo	Timing	Structure	Key	Harm.	Lyrics	Noise
Remaster	⋆							
Instrumental	⋆						⋆	⋆
Live	⋆	⋆	⋆					⋆
Acoustic	⋆	⋆	⋆		⋆	⋆		⋆
Demo	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
Medley	⋆	⋆	⋆	⋆	⋆			⋆
Remix	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
Quotation	⋆			⋆				⋆

done in areas such as query-by-humming systems, content-based music retrieval, genre classification, or audio fingerprinting, is relevant for addressing cover song similarity.

Many ideas for cover song identification systems come from the symbolic domain [45, 56, 67, 81], and query-by-humming systems [12] are paradigmatic examples. In query-by-humming systems, the user sings or hums a melody and the system searches for matches in a musical database. This query-by-example situation is parallel to retrieving cover songs from a database. In fact, many of the note encoding or alignment techniques employed in query-by-humming systems could be useful in future approaches for cover song identification. However, the kind of musical information that query-by-humming systems manage is symbolic (usually MIDI files), and the query, as well as the music material, must be transcribed into the symbolic domain. Unfortunately, transcription systems of this kind do not yet achieve a significantly high accuracy on real-world audio music signals. Current state-of-the-art algorithms yield overall accuracies around 75%[10], even for melody estimation[11], indicating that there is still much room for improvement in these areas. Consequently, we argue that research in the symbolic domain cannot be directly applied to audio domain cover song similarity systems without incurring several estimation errors in the first processing stages of these. These errors, in turn, may have dramatic consequences in final system's accuracy.

Content-based music retrieval is organized around use cases which define a type of query, the sense of match, and the form of the output [10, 18]. The sense of match implies different degrees of specificity: it can be exact, retrieving music with specific content, or approximate, retrieving near neighbors in a musical space where proximity encodes different senses of musical similarity [10]. One prototypical use case is genre classification [70]. In this case, one generally tries to group songs according to a commercially or culturally established label, the genre, where

[10] http://www.music-ir.org/mirex/2008/index.php/Multiple_Fundamental_Frequency_Estimation_&_Tracking_Results

[11] http://www.music-ir.org/mirex/2008/index.php/Audio_Melody_Extraction_Results

certain characteristics might be more or less the same but many others might radically change (category-based song grouping). Therefore, genre classification is considered to have a low match specificity [10]. On the other hand, audio fingerprinting [7] is an example of a task with a highly specific match. This essentially consists in identifying a particular performance of a concrete song (exact duplicate detection). In contrast to many prototypical use cases, cover song identification is representative of an intermediate specificity region [10]. It goes beyond audio fingerprinting in the sense that it tries to approximate duplicate detection while allowing many musical facets to change. In addition, it is more specific than genre classification in the sense that it goes beyond timbral similarity to include the important idea that musical works retain their identity notwithstanding variations in many musical dimensions [19].

It must be noted that many studies approach the aforementioned intermediate match specificity. This is the case, for instance, of many audio fingerprinting algorithms using tonality-based descriptors instead of the more routinely employed timbral ones (e.g. [8, 51, 66, 84]). These approaches can also be named with terms such as audio identification, audio matching, or simply, polyphonic audio retrieval. The adoption of tonal features adds some degrees of invariance (timbre, noise) to audio fingerprinting algorithms which are, by nature, invariant with respect to song structure changes. In spite of that, many of them might still have a low recall for cover versions. This could be due to an excessively coarse feature quantization [66], and to the lack of other desirable degrees of invariance to known musical changes like tempo variations or key transpositions [76].

Like recent audio identification algorithms, many other systems derived from the genre classification task or from traditional music similarity approaches may also fall into the aforementioned intermediate specificity region. These, in general, differ from traditional systems of their kind in the sense that they also incorporate tonal information (e.g. [48, 61, 82, 90]). However, these systems might also fail in achieving invariance to key or tempo modifications. In general, they do not consider full sequences of musical events, but just statistical summarizations of them, which might blur/distort valuable information for assessing the similarity between cover songs.

Because of the large volume of existing work it is impossible to cover every top in this area. We focus on algorithms designed for cover song identification, that, in addition, include several modules explicitly designed to achieve invariance to characteristic musical changes among versions[12].

2 Approaches

The standard approach to measuring similarity between cover songs is essentially to exploit music facets shared between them. Since several important characteristics

[12] Even considering this criteria, it is difficult to present the complete list of methods and alternatives. We apologize for possible omissions/errors and, in any case, we assert that these have not been intentional.

are subject to variation (timbre, key, harmonization, tempo, timing, structure, and so forth, Section 1.3), cover song identification systems must be robust against these variations.

Extracted descriptors are often in charge of overcoming the majority of musical changes among covers, but special emphasis is put on achieving tempo, key, or structure invariance, as these are very frequent changes that are not usually managed by extracted descriptors themselves. Therefore, one can group the elements of existing cover song identification systems into four basic functional blocks: feature extraction, key invariance, tempo invariance, and structure invariance. An extra block can be considered at the end of the chain for the final similarity measure used (figure 1 illustrates these blocks). A summary table for several state-of-the-art approaches, and the different strategies they follow in each functional block, is provided at the end of the present section (table 2).

2.1 Feature Extraction

In general, we can assume that different versions of the same piece mostly preserve the main melodic line and/or the harmonic progression, regardless of its main key. For this reason, tonal or harmonic content is a mid-level characteristic that should be considered to robustly identify covers.

The term tonality is commonly used to denote a system of relationships between a series of pitches, which can form melodies and harmonies, having a tonic (or central pitch class) as its most important (or stable) element [42]. In its broadest possible sense, this term refers to the arrangements of pitch phenomena. Tonality is ubiquitous in Western music, and most listeners, either musically trained or not, can identify the most stable pitch while listening to tonal music [11]. Furthermore, this process is continuous and remains active throughout the sequential listening experience [72].

A tonal sequence can be understood, in a broad sense, as a sequentially-played series of different note combinations. These notes can be unique for each time slot (a melody) or can be played jointly with others (chord or harmonic progressions). From a MIR point of view, clear evidence about the importance of tonal sequences for music similarity and retrieval exists [9, 22, 34]. In fact, almost all cover song identification algorithms exploit tonal sequence representations extracted from the raw audio signals: they either estimate the main melody, the chord sequence, or the harmonic progression. Only early systems, which, e.g., work with the audio signal's energy or with spectral-based timbral features, are an exception [25, 89].

Melody is a salient musical descriptor of a piece of music [73] and, therefore, several cover song identification systems use melody representations as a main descriptor [49, 50, 68, 78, 79]. As a first processing step, these systems need to extract the predominant melody from the raw audio signal [62]. Melody extraction is strongly related to pitch tracking, which itself has a long and continuing history [13]. However, in the context of complex mixtures, the pitch tracking problem becomes further complicated because, although multiple pitches may be present at the same time, at

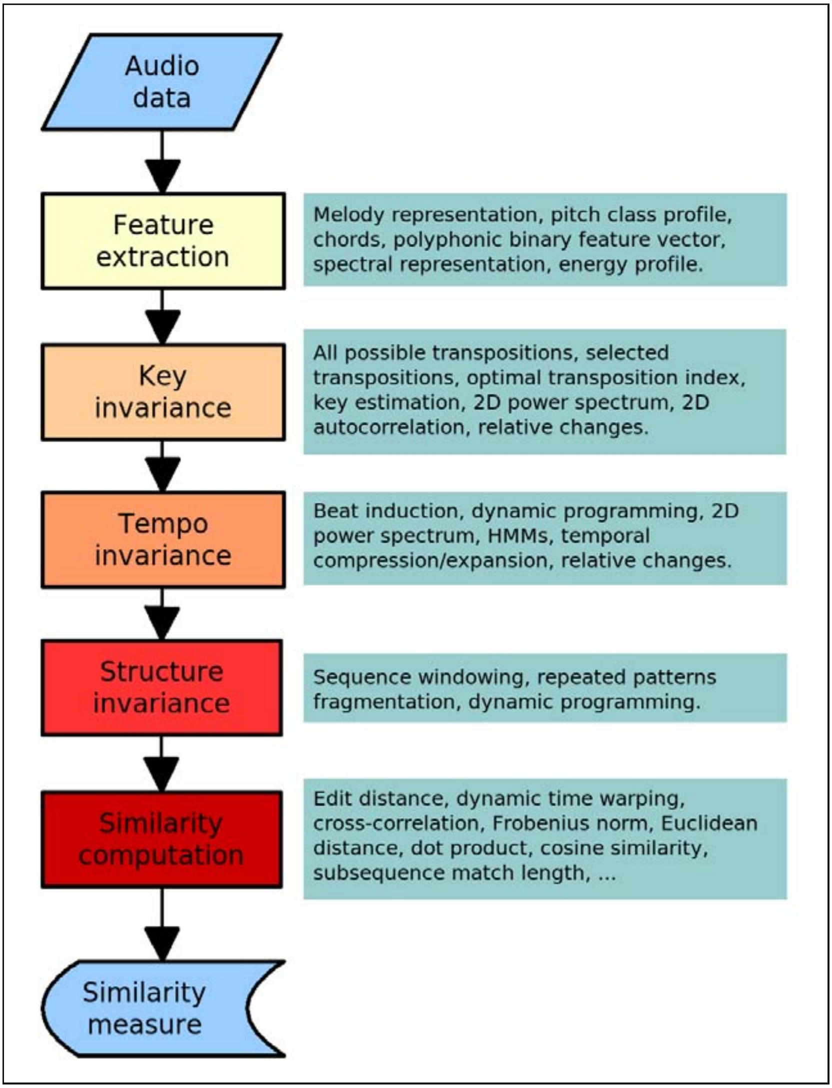
Audio data
Feature extraction
Melody representation, pitch class profile, chords, polyphonic binary feature vector, spectral representation, energy profile.
Key invariance
All possible transpositions, selected transpositions, optimal transposition index, key estimation, 2D power spectrum, 2D autocorrelation, relative changes.
Tempo invariance
Beat induction, dynamic programming, 2D power spectrum, HMMs, temporal compression/expansion, relative changes.
Structure invariance
Sequence windowing, repeated patterns fragmentation, dynamic programming.
Similarity computation
Edit distance, dynamic time warping, cross-correlation, Frobenius norm, Euclidean distance, dot product, cosine similarity, subsequence match length, ...
Similarity measure

Fig. 1 Generic block diagram for cover song identification systems

most just one of them will be the melody. This and many other facets [62] make melody extraction from real-world audio signals a difficult task (see Section 1.4). To refine the obtained representation, cover detection systems usually need to combine a melody extractor with a voice/non-voice detector and other post-processing modules in order to achieve a more reliable representation [68, 78, 79]. Another possibility is to generate a so-called "mid-level" representation for these melodies [49, 50], where the emphasis is not only put on melody extraction, but also on the feasibility to describe audio in a way that facilitates retrieval.

Alternatively, cover song similarity can be assessed by harmonic, rather than melodic, sequences using so-called chroma features or pitch class profiles (PCP) [26, 27, 63, 82]. These mid-level features might provide a more complete, reliable, and straightforward representation than, e.g., melody estimation, as they do not need to tackle the pitch selection and tracking issues outlined above. PCP features are derived from the energy found within a given frequency range (usually from 50 to 5000 Hz) in short-time spectral representations (typically 100 msec) of audio signals extracted on a frame-by-frame basis. This energy is usually collapsed into a 12-bin octave-independent histogram representing the relative intensity of each of the 12 semitones of an equal-tempered chromatic scale. Reliable PCP features should, ideally, (a) represent the pitch class distribution of both monophonic and polyphonic signals, (b) consider the presence of harmonic frequencies, (c) be robust to noise and non-tonal sounds, (d) be independent of timbre and played instrument, (e) be independent of loudness and dynamics, and (f) be independent of tuning, so that the reference frequency can be different from the standard A 440 Hz [27]. This degree of invariance with respect to several musical characteristics make PCP features very attractive for cover song identification systems. Hence, the majority of systems use a PCP-based feature as primary source of information [20, 21, 23, 28, 29, 36, 37, 38, 39, 40, 41, 55, 53, 76, 74].

An interesting variation of using PCP features for characterizing cover song similarity is proposed in [9]. In this work, PCP sequences are collapsed into string sequences using vector quantization, i.e. summarizing several features vectors by a close representative, done via the K-means algorithm [88] (8, 16, 32, or 64 symbols). In [55], vector quantization is performed by computing binary PCP feature vector components in such a way that, with 12 dimensional feature vectors, a codebook of $2^{12} = 4096$ symbols is generated (so-called polyphonic binary feature vectors). Sometimes, the lack of interpretability of the produced sequences and symbols makes the addition of some musical knowledge to these systems rather difficult. This issue is further studied in [41] where, instead of quantizing in a totally unsupervised way, a codebook of PCP features based on musical knowledge (with a size of 793 symbols) is generated. In general, vector quantization, indexing, and hashing techniques, result in highly efficient algorithms for music retrieval [8, 41, 55, 66], even though their accuracy has never been formally assessed for the specific cover song identification task. It would be very interesting to see how these systems perform on a benchmark cover song training set (e.g. MIREX [18]) in comparison to specifically designed approaches. More concretely, it is still an issue if PCP quantization strongly degrades cover song retrieval. Some preliminary results suggest that this is the case [66].

Instead of quantizing PCP features, one can use chord or key template sequences for computing cover song similarity [2, 4, 35, 43]. Estimating chord sequences from audio data has been a very active research area in recent years [5, 44, 60, 77]. The common process for chord estimation consists of two steps: pre-processing the audio into a feature vector representation (usually a PCP feature), and approximating the most likely chord sequence from these vectors (usually done via template-matching or expectation-maximization trained Hidden Markov Models

[64]). Usually, 24 chords are used (12 major and 12 minor), although some studies incorporate more complex chord types, such as 7^{th}, 9^{th}, augmented, and diminished chords [26, 32]. This way, the obtained strings have a straightforward musical interpretation. However, chord-based tonal sequence representations might be too coarse for the task at hand if one considers previously mentioned PCP codebook sizes, and might be also error-prone.

2.2 Key Invariance

As mentioned in section 1.3, cover songs may be transposed to different keys. Transposed versions are equivalent to most listeners, as pitches are perceived relative to each other rather than in absolute categories [16]. In spite of being a common change between versions, some systems do not explicitly consider transpositions. This is the case for systems that do not specifically focus on cover songs, or that do not use a tonal representation [25, 35, 53, 89].

Transposition is reflected as a ring-shift with respect to the "pitch axis" of the feature representation. Several strategies can be followed to tackle transposition, and their suitability may depend on the chosen feature representation. In general, transposition invariance can be achieved by relative feature encoding, by key estimation, by shift-invariant transformations, or by applying different transpositions.

The most straightforward way to achieve key invariance is to test all possible feature transpositions [21, 23, 36, 38, 39, 41, 50, 55]. In the case of an octave-independent representation, this implies the computation of a similarity measure for all possible circular (or ring) shifts in the pitch axis for each song. This strategy usually guarantees a maximal retrieval accuracy [75] but, on the other hand, either the time or the size (or both) of the database to search in increases.

Recently, some speeding-up approaches for this process have been presented [75, 76]. Given a tonal representation for two songs, these algorithms basically compute the most probable relative transpositions given an overall representation of the tonal content of each song (the so-called optimal transposition index) [20, 74, 76]. This process is very fast since this overall representation can be, e.g., a simple averaging of the PCP features over the whole sequence, and can be calculated off-line. Finally, only the K most probable shifts are chosen. Further evaluation suggests that, for 12 bin PCP-based representations, a near-optimal accuracy can be reached with just two shifts [75], thus reducing six times the computational load. Some systems do not follow these strategy and predefine a certain number of transpositions to compute. These can be chosen arbitrarily [78, 79], or based on some musical and empirical knowledge [4]. Decisions of this kind are very specific for each system.

An alternative approach is to off-line estimate the main key of the song and then apply transposition accordingly [28, 29, 49]. In this case, errors propagate faster and can dramatically worsen retrieval accuracy [75, 76] (e.g. if the key for the original song is not correctly estimated, no covers will be retrieved as they might have been estimated in the correct one). However, it must be noted that a similar procedure to choosing the K most probable transpositions could be employed.

If a symbolic representation such as chords is used, one can further modify it in order to just describe relative chord changes. This way, a key-independent feature sequence is obtained [2, 43, 68]. This idea, which is grounded in existing research on symbolic music processing [12, 45, 56, 67, 81], has been recently extended to PCP sequences [40, 38] by using the concept of optimal (or minimizing) transposition indices [52, 76].

A very interesting approach to achieve transposition invariance is to use a 2D power spectrum [50] or a 2D autocorrelation function [37]. Autocorrelation is a well-known operator for converting signals into a delay or shift-invariant representation [58]. Therefore, the power spectrum (or power spectral density), which is formally defined as the Fourier transform of the autocorrelation, is also shift-invariant. Other 2D transforms (e.g. from image processing) could be also used, specially shift-invariant operators derived from higher-order spectra [33].

2.3 Tempo Invariance

Different renditions of the same piece may vary in the speed they have been played, and any descriptor sequence extracted in a frame-by-frame basis from these performances will reflect this variation. For instance, in case of doubling the tempo, frames $i, i+1, i+2, i+3$ might correspond to frames $j, j, j+1, j+1$, respectively. Consequently, extracted sequences cannot be directly compared. Some cover song identification systems fail to include a specific module to tackle tempo fluctuations [2, 38, 39, 90, 91]. The majority of these systems generally focus on retrieval efficiency and treat descriptor sequences as statistical random variables. Thus, they throw away much of the sequential information that a given representation can provide (e.g. a representation consisting of a 4 symbol pattern like ABABCD, would yield the same values as AABBCD, ABCABD, etc., which is indeed a misleading oversimplification of the original data).

In case of having a symbolic descriptor sequence (e.g. the melody), one can encode it by considering the ratio of durations between two consecutive notes [68]. This strategy is employed in query-by-humming systems [12] and, combined with relative pitch encoding (section 2.3), leads to a representation that is key and tempo independent. However, for the reasons outlined in section 2.1, extracting a symbolic descriptor sequence is not straightforward and may lead to important estimation errors. Therefore, one needs to look at alternative tempo-invariance strategies.

One way of achieving tempo invariance is to estimate the tempo and then aggregate the information contained within comparable units of time. In this manner, the usual strategy is to estimate the beat [30] and then aggregate the descriptor information corresponding to the same beat. This can be done independently of the descriptor used. Some cover song identification systems based on a PCP [23, 55] or a melodic [49, 50] representation use this strategy, and extensions with chords or other types of information could be easily devised. If the beat does not provide enough temporal resolution, a finer representation (e.g. half-beat, quarter-beat, etc.) might be employed [21]. However, some studies suggest that systems using beat-averaging strategies can be outperformed by others, specially the ones employing dynamic programming [4, 76].

An alternative to beat induction is doing temporal compression/expansion [41, 53]. This straightforward strategy consists in re-sampling the signal into several musically plausible compressed/expanded versions and then comparing all of them in order to empirically discover the correct re-sampling.

Another interesting way to achieve tempo independence is again the 2D power spectrum or the 2D autocorrelation function [36, 37, 50]. These functions are usually designed for achieving both tempo as well as key independence, but 1D versions can also be designed (section 2.2).

If one wants to perform direct frame to frame comparison, a sequence alignment/similarity algorithm must be used to determine frame to frame correspondence between two song's representations. Several alignment algorithms for MIR have been proposed (e.g. [1, 15, 52]) which, sometimes, derive from general string and sequence alignment/similarity algorithms [31, 65, 69]. In cover song identification, dynamic programming [31] is a routinely employed technique for aligning two representations and automatically discovering their local correspondences [4, 20, 25, 28, 29, 35, 43, 49, 55, 74, 76, 78, 79, 89]. Overall, one iteratively constructs a cumulative distance matrix by considering the optimal alignment paths that can be derived by following some neighboring constraints (or patterns) [54, 65]. These neighboring constraints determine the allowed local temporal deviations and they have been evidenced to be an important parameter in the final system's accuracy [54, 76]. One might hypothesize that this importance relies on the ability to track local timing variations between small parts of the performance (section 1.3). For cover song identification, dynamic programming algorithms have been found to outperform beat induction strategies [4, 76]. The most typical algorithms for dynamic programming alignment/similarity are dynamic time warping algorithms [65, 69] and edit distance variants [31]. Their main drawback is that they are computationally expensive (i.e., quadratic in the length of the song representations), but several fast implementations may be derived [31, 56, 83].

2.4 Structure Invariance

The difficulties that a different song structure may pose in the computation of a cover song similarity measure are very often neglected. However, this has been demonstrated to be a key factor [76] and actually, recent cover song identification systems thoughtfully consider this aspect, especially many of the best-performing ones[13].

A classic approach to structure invariance consists in summarizing a song into its most repeated or representative parts [29, 49]. In this case, song structure analysis [57] is performed in order to segment sections from the song's representation used. Usually, the most repetitive patterns are chosen and the remaining patterns are disregarded. This strategy might be prone to errors since structure segmentation algorithms still have much room for improvement [57]. Furthermore, sometimes the

[13] For accuracies please see section 3 and references therein.

most identifiable or salient segment for a song is not the most repeated one, but the introduction, the bridge, and so forth.

Some dynamic programming algorithms deal with song structure changes. These are basically the so-called local alignment algorithms [31], and have been successfully applied to the task of cover song identification [20, 74, 76, 89]. These systems solely consider the best subsequence alignment found between two song's representation for similarity assessment, what has been evidenced to yield very satisfactory results [76].

However, the most common strategy for achieving structure invariance consists in windowing the descriptors representation (sequence windowing) [41, 50, 53, 55]. This windowing can be performed with a small hop size in order to faithfully represent any possible offset in the representations. This hop size has not been found to be a critical parameter for accuracy, as near-optimal values are found for a considerable hop size range [50]. Sequence windowing is also used by many audio fingerprinting algorithms using tonality-based descriptors [8, 51, 66], and it is usually computationally less expensive than dynamic programming techniques for achieving structural invariance.

2.5 Similarity Computation

The final objective of a cover song identification system is, given a query, to retrieve a list of cover songs from a music collection. This list is usually ranked according to some similarity measure so that first songs are the most similar to the query. Therefore, cover song identification systems output a similarity (or dissimilarity[14]) measure between pairs of songs. This similarity measure operates on the obtained representation after feature extraction, key invariance, tempo invariance, and structure invariance modules.

Common dynamic programming techniques used for achieving tempo invariance (section 2.3) already provide a similarity measure as an output [31, 65, 69]. Accordingly, the majority of the cover song identification systems following a dynamic programming approach use the similarity measure these approaches provide. This is the case for systems using edit distances [4, 68] or dynamic time warping algorithms [25, 28, 29, 35, 43, 78, 79]. These similarity measures usually contain an implicit normalization depending on the representation's lengths, which can generate some conflicts with versions of very different durations. In the case of local alignment dynamic programming techniques (section 2.4), the similarity measure usually corresponds to the length of the found subsequence match [20, 55, 74, 76, 89].

Conventional similarity measures like cross-correlation [21, 23, 49], the Frobenius norm [36], the Euclidean distance [37, 50], or the dot product [38, 39, 41, 53] are also used. They are sometimes normalized depending on compared representation's lengths. In the case of adopting a sequence windowing strategy for dealing

[14] For the sake of generality, we use the term similarity to refer to both the similarity and the dissimilarity. In general, a distance measure can also be considered a dissimilarity measure, which, in turn, can be converted to a similarity measure.

Table 2 Cover song identification methods and their ways to overcome departures from the ”canonical” song. A blank space denotes no specific treatment for them. Abbreviations for extracted features are PBFV for polyphonic binary feature vector, and PCP for pitch class profile. Abbreviation for key invariance is OTI for optimal transposition index. Abbreviations for tempo invariance are DP for dynamic programming, and HMM for Hidden Markov Models. Abbreviations for similarity computation are DTW for dynamic time warping, MLSS for most likely sequence of states, and NCD for normalized compression distance.

Reference(s)	Extracted feature	Key invariance	Tempo invariance	Structure invariance	Similarity computation
Ahonen & Lemstrom [2]	Chords	Relative changes			NCD
Bello [4]	Chords	*K* transpositions	DP		Edit distance
Egorov & Linetsky [20]	PCP	OTI	DP	DP	Match length
Ellis et al. [21, 23]	PCP	All transpositions	Beat		Cross-correlation
Foote [25]	Energy + Spectral		DP		DTW
Gómez & Herrera [28]	PCP	Key estimation	DP		DTW
Gómez et al. [29]	PCP	Key estimation	DP	Repeated patterns	DTW
Izmirli [35]	Key templates		DP		DTW
Jensen et al. [36]	PCP	All transpositions	Fourier transform		Frobenius norm
Jensen et al. [37]	PCP	2D autocorrelation	2D autocorrelation		Euclidean distance
Kim et al.[38, 39]	PCP + Delta PCP	All transpositions			Dot product
Kim & Perelstein [40]	PCP	Relative changes	HMM		MLSS
Kurth & Muller [41]	PCP	All transpositions	Temporal comp./exp.	Sequence windowing	Dot product
Lee [43]	Chords	Key estimation	DP		DTW
Marolt [49]	Melodic	Key estimation	DP	Repeated patterns	Cross-correlation
Marolt [50]	Melodic	2D spectrum	Beat + 2D spectrum	Sequence windowing	Euclidean distance
Müller et al. [53]	PCP		Temporal comp./exp.	Sequence windowing	Dot product
Nagano et al. [55]	PBFV	All transpositions	Beat + DP	Seq. windowing + DP	Match length
Sailer & Dressler [68]	Melodic		Relative changes		Edit distance
Serrà et al. [74, 76]	PCP	OTIs	DP	DP	Match length
Tsai et al. [78, 79]	Melodic	*K* transpositions	DP		DTW
Yang [89]	Spectral		DP	Linearity filtering	Match length

with structure changes (section 2.4), these similarity measures are usually combined with multiple post-processing steps such as threshold definition [41, 53, 50], TF-IDF[15] weights [50], or mismatch ratios [41]. Less conventional similarity measures include the normalized compression distance [2], and the Hidden Markov Model-based most likely sequence of states [40]. In table 2 we show a summary of all the outlined approaches and their strategies for overcoming musical changes among cover versions and for similarity computation.

3 Evaluation

The evaluation of cover song identification and similarity systems is a complex task, and it is difficult to find in the literature a common methodology for that. The only existing attempt to compare version identification systems is found in the Music Information Retrieval Evaluation eXchange (MIREX[16]) initiative [18, 19]. Nevertheless, the MIREX framework only provides an overall accuracy of each system. A valuable improvement would be to implement independent evaluations for the different processes involved (feature extraction, similarity computation, etc.), in order to analyze their contributions to the global system behavior.

The evaluation of cover song identification systems is usually set up as a typical information retrieval "query and answer" task [3], where one submits a query song and the system returns a ranked set (or list) of answers retrieved from a given collection [19]. Then, the main purpose of the evaluation process is to assess how precise the retrieved set is. We discuss two important issues regarding the evaluation of cover song retrieval systems: the evaluation measures and the music material used.

3.1 Evaluation Measures

A referential evaluation measure might be the mean of average precision (MAP). This measure is routinely employed in various information retrieval disciplines [3, 47, 86] and some works on cover song identification have recently started reporting results based on it [2, 20, 74]. In addition, it has been also used to evaluate the cover song identification task in the MIREX [19].

Although MIREX defines some evaluation measures, in the literature there is no agreement on which one to use. Therefore, in addition to MAP, several other measures have been proposed. These include the R-Precision (R-Prec, [4, 35]), variants of Precision or Recall at different rank levels (P@X, R@X, [21, 23, 25, 36, 37, 38, 39, 41, 78, 79, 89]), the average of Precision and Recall (Avg PR, [55]), and the F-measure (Fmeas, [28, 29, 76]).

[15] The TF-IDF weight (term frequency-inverse document frequency) is a weight often used in information retrieval and text mining. For more details we refer to [3].

[16] `http://www.music-ir.org/mirexwiki/index.php/Main_Page`

3.2 Music Material: Genre, Variability, and Size Issues

One relevant issue when dealing with evaluation is the considered music material. Both the complexity of the problem and the selected approach largely depend on the studied music collection and the types of versions we want to identify, which might range from remastered tracks, to radically different songs (section 1.2). In this sense, it is very difficult to compare two systems evaluated in different conditions and designed to solve different problems. Some works solely analyze classical music [35, 38, 39, 41, 53], and it is the case that all of them obtain very high accuracies. However, classical music versions might not present strong timbral, structural, or tempo variations. Therefore, one might hypothesize that, when only classical music is considered, the complexity of the cover song identification task decreases. Other works use a more variated style distribution in their music collections [2, 4, 21, 23, 36, 37, 49, 50] but many times it is still unclear which types of versions are used. These are usually mixed and may include remastered tracks (which might be easier to detect), medleys (where invariance towards song structure changes may be a central aspect), demos (with substantial variations with respect to the finally released song), remixes, or quotations (which might constitute the most challenging scenario due to their potentially low duration and distorted harmonicity). The MIREX music collection is meant to include a wide variety of genres (e.g. classical, jazz, gospel, rock, folk-rock, etc.), and a sufficient variety of styles and orchestrations [19]. However, the types of covers that are present in the MIREX collection are unknown[17]. In our view, a big variety in genres and types of covers is the only way to ensure the general applicability of the method being developed.

Apart from the qualitative aspects of the considered music material, one should also care with the quantitative aspects of it. The total amount of songs and the distribution of these can strongly influence final accuracy values. To study this influence, one can decompose a music collection into cover sets (i.e. each original song is assigned to a separate cover set). Then, their cardinality (number of covers per set, i.e., the number of covers for each original song) becomes an important parameter. A simple test was performed with the system described in [74] in order to assess the influence of these two parameters (number of cover sets, and their cardinality) on the final system's accuracy. Based on a collection of 2135 cover songs, 30 random selections of songs were carried out according to the aforementioned parameters. Then, average MAP for all runs was computed and plotted (figure 2). We can see that considering less than 50 cover sets or even just a cardinality of 2 yields unrealistically high results, while higher values for these two parameters at the same time all fall in a stable accuracy region[18]. This effect can also be seen if we plot the standard deviations of the evaluation measure across all runs (figure 3). Finally, it can be observed that using less than 50 cover sets introduces a high variability in the

[17] As the underlying datasets are not disclosed, information of this kind is unavailable.

[18] It is not the aim of the experiment to provide explicit accuracy values. Instead, we aim at illustrating the effects that different configurations of the music collection might have for final system's accuracy.

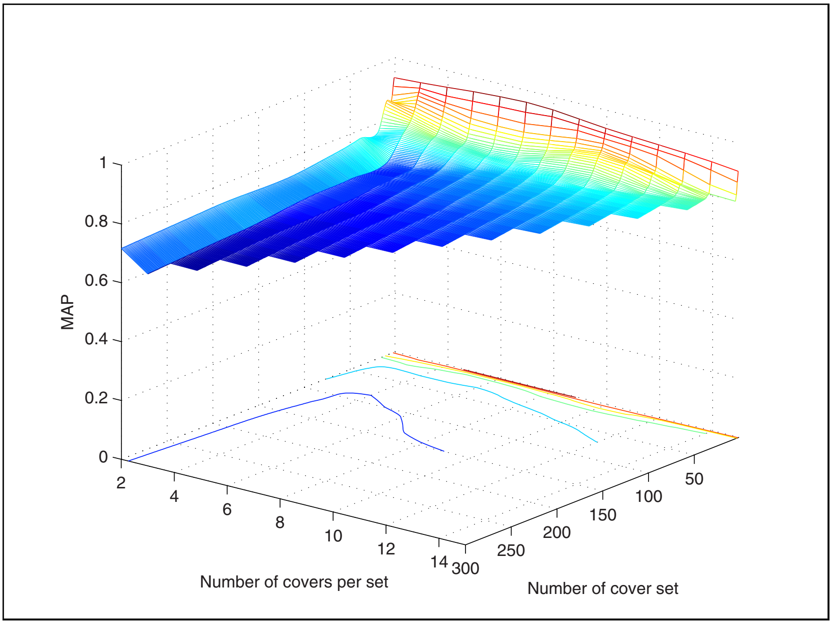

Fig. 2 Accuracy of a cover song identification system depending on the number of cover sets, and the number of covers per set

evaluated accuracy, which might then depend on the chosen subset. This variability becomes lower as the number of cover sets and their cardinality increase.

With the small experiment above, we can see that an insufficient size of the music collection could potentially lead to abnormally high accuracies, as well as to parameter over fitting (in the case of requiring a training procedure for some of them). Unfortunately, many reported studies use less than 50 cover sets [25, 28, 29, 35, 55, 78, 79]. Therefore, one cannot be confident about the reported accuracies. This could even happen with the so-called *covers80* cover song dataset[19] (a freely available dataset that many researchers use to test system's accuracy and to tune their parameters [2, 21, 23, 36, 37]), which is composed of 80 cover sets with a cardinality of 2.

In case when the evaluation dataset is not large enough, one may try to compensate the artifacts this might produce by adding so-called 'noise' or 'control' songs [4, 20, 49, 50]. The inclusion of these songs in the retrieval collection might provide an extra dose of difficulty to the task, as the probability of getting relevant items within the first ranked elements becomes then very low [19]. This approach is also followed within the MIREX framework. Therefore, test data is composed of thirty cover sets, each one consisting of eleven different versions. Accordingly, the total cover song collection contains 330 songs. In order to make the detection task more difficult, 670 individual songs, i.e., cover sets of cardinality 1, are added [19].

[19] `http://labrosa.ee.columbia.edu/projects/coversongs/covers80`

Table 3 Cover song identification methods and their evaluation strategies. Accuracies (including MIREX) correspond to best result achieved. Blank space, ‘ˆ’, and ‘˜’ denote unknown, approximate, and average values, respectively. Legend for genres is (B) blues, (C) classical, (CO) country, (E) electronic, (J) jazz, (HH) hip-hop, (M) metal, (P) pop, (R) rock, and (W) world. Legend for types of covers is (A) acoustic, (DE) demo, (DU) duet, (I) instrumental, (L) live, (M) medley, (RR) remaster, (RX) remix, and (Q) quotation.

Reference(s)	Cover sets	Covers per set	Total	Musical styles	Types of covers	Evaluation measure	Accuracy	MIREX MAP
Ahonen & Lemstrom [2]	80	2	160	B, CO, M, P, R	A, DU, I, L	MAP	0.18	
Bello [4]	36	4.4˜	3208	P, R,	L,	R-Prec	0.25	0.27
Egorov & Linetsky [20]	30	11	1000	C, CO, E, HH, MT, P, R	A, DU, I, L, RR,	MAP	0.72	0.55
Ellis et al. [21, 23]	80	2	160	B, CO, M, P, R	A, DU, I, L	P@1	0.68	0.33
Foote [25]	28		82	C, P	A, I, L,	P@3	0.80	
Gómez & Herrera [28]	30	3.1˜	90			Fmeas	0.39	
Gómez et al. [29]	30	3.1˜	90			Fmeas	0.41	
Izmirli [35]	12		125	C		R-Prec	0.93	
Jensen et al. [36]	80	2	160	B, CO, M, P, R	A, DU, I, L	P@1	0.38	0.24
Jensen et al. [37]	80	2	160	B, CO, M, P, R	A, DU, I, L	P@1	0.48	0.23
Kim et al.[38, 39]	1000ˆ	2˜	2000	C		P@1	0.79	
Kim & Perelstein [40]								0.06
Kurth & Muller [41]			1167	C		R@1	0.97	
Lee [43]								0.13
Marolt [49]	8	4.5˜	1820	P, R		P@5	0.22	
Marolt [50]	34	4.3˜	2424	P, R		MAP	0.40	
Müller et al. [53]			1167	C		P@15	0.93	
Nagano et al. [55]	8	27˜ˆ	216		L	Avg PR	0.89	
Sailer & Dressler [68]								0.07ˆ
Serrà et al. [74, 76]	525	4.1˜	2135	B, C, CO, E, J, M, P, R, W	A, DE, DU, I, L, M, RX, Q	MAP	0.66	0.66
Tsai et al. [78, 79]	47	2	794			P@1	0.77	
Yang [89]			120	C, P, R		P@2	0.99	

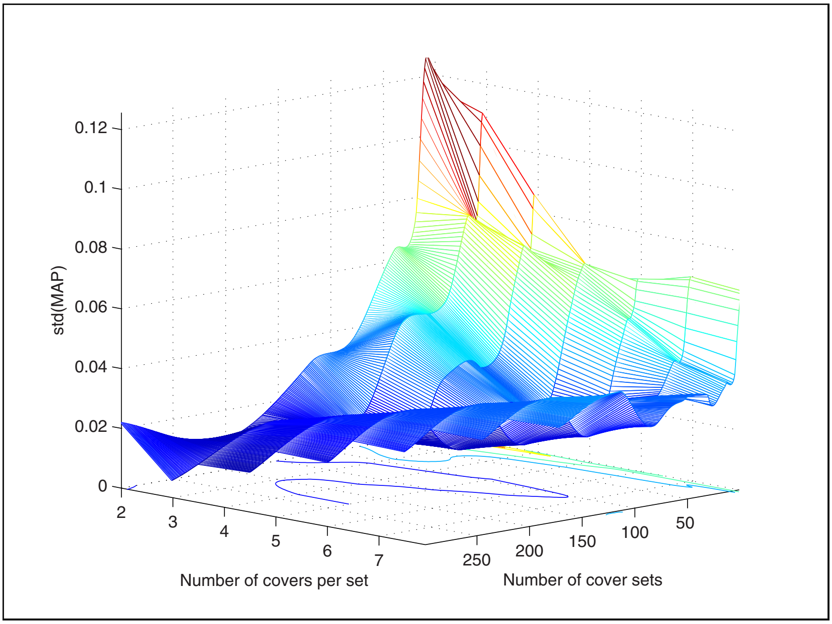

Fig. 3 Standard deviation of a cover song identification system's accuracy depending on the number of cover sets, and the number of covers per set

As a corollary, we could hypothesize that the bigger and more varied the music collection is, the more similar the out-of-sample results (and therefore better scalability) we shall obtain. In addition, one should stress that the usage of an homogeneous and small music collection, apart from leading to abnormal accuracies, could also lead to incorrect parameter estimates. In table 3 we show a summary of the evaluation strategies and accuracies reported by the cover song identification systems outlined in section 2.

4 Concluding Remarks

We have summarized here the work done for addressing the problem of automatic cover song identification. Even though different approaches have been tried, it seems quite clear that a well-crafted system has to be able to exploit tonal, temporal, and structural invariant representations of music. We have also learnt that there are methodological issues to be considered when building music collections used as ground-truth for developing and evaluating cover identification systems.

Once we have concluded this exhaustive overview, some conceptual open issues can be remarked. Even though the main invariances to be computed correspond to tonal and rhythm information, we still ignore the role (if any) of timbre, vocal features, or melodic similarity. Timbre similarity, including vocal similarity for sung music, could have some impact for identifying those covers intended to be close matches to a given query. In other situations this type of similarity would be misleading, though. Finding an automated way for deciding on that is still an open research issue.

In order to determine a similarity measure between cover songs, the usual approach pays attention solely to the musical facets that are shared among them (section 2). This makes sense if we suppose that these changes do not affect the similarity between covers. For instance, if two songs are covers and have the same timbre characteristics, and a third song is also a cover but does not exhibit the same timbre, they will score the same similarity. This commonality-based sense of similarity contrasts with the feature contrast similarity model [80], wherein similarity is determined by both common and distinctive features of the objects being compared. Future works approaching cover song similarity in a stricter sense might benefit from considering also differences between them, so that, in the previous example, the third cover is less similar than the two first ones.

Determining cover song similarity in a stricter sense would have some practical consequences and would be a useful feature for music retrieval systems. Therefore, depending on the goals of the listeners, different degrees of similarity could be required. Here we have a new scenario where the ill-defined but typical music similarity problem needs to be addressed. Research reported in this chapter could provide reasonable similarity metrics for this, but preservation of timbral and structural features would be required in addition, in order to score high in similarity with respect to the query song.

Another avenue for research is that of detecting musical quotations. In classical music, there is a long tradition of composers citing phrases or motives from other composers (e.g. Alban Berg quoting Bach's chorale *Es ist genug* in his *Violin Concerto*, or Richard Strauss quoting Beethoven's *Eroica symphony* in his 'Metamorphosen for 23 solo strings'). In popular music there are also plenty of quotations (e.g. The Beatles' ending section of *All you need is love* quotes the French anthem *La Marseillaise* and Glen Miller's *In the mood*, or Madonna's *Hung up* quoting Abba's *Gimme, Gimme, Gimme*), and even modern electronic genres massively borrow loops and excerpts from any existing recording. As the quoted sections are usually of short duration, special adaptations of the reviewed algorithms would be required to detect them. In addition to facilitating law enforcement procedures, linking this way diverse musical works opens new interesting ways for navigating across huge music collections.

Beyond many conceptual open issues, there are still some technical aspects that deserve effort to improve the efficiency of a system. First, perfecting a music processing system requires careful examination and analysis of errors. When errors are patterned they can reveal specific deficiencies or shortcomings in the algorithm. We are still lacking of that kind of in-depth analysis. Second, rigorously evaluating a cover song similarity metric would require the ground truth songs to be categorized according to the musical facets involved (section 1.3) and, maybe, according to the cover song category they belong to (section 1.2). Third, achieving a robust, scalable, and efficient method is still an issue. It is outstanding that systems achieving the highest accuracies are quite computationally expensive, and that fast retrieval systems fail in recognizing many of the cover songs a music collection might contain (section 2.1). We hypothesize that there exists a trade-off between system's accuracy and efficiency. However, we believe that these and many other technical as well as conceptual issues might be overcome in next years.

Acknowledgments

The authors want to thank their colleagues and staff at the Music Technology Group (UPF) for their support and work. This research has been partially funded by the EU-IP project PHAROS[20] (IST-2006-045035) and the e-Content Plus project VARIAZIONI[21] (ECP-2005-CULT-038264).

References

1. Adams, N.H., Bartsch, N.A., Shifrin, J.B., Wakefield, G.H.: Time series alignment for music information retrieval. In: Int. Symp. on Music Information Retrieval (ISMIR), pp. 303–310 (2004)
2. Ahonen, T.E., Lemstrom, K.: Identifying cover songs using normalized compression distance. In: Int. Workshop on Machine Learning and Music, MML (July 2008)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press Books, New York (1999)
4. Bello, J.P.: Audio-based cover song retrieval using approximate chord sequences: testing shifts, gaps, swaps and beats. In: Int. Symp. on Music Information Retrieval (ISMIR), September 2007, pp. 239–244 (2007)
5. Bello, J.P., Pickens, J.: A robust mid-level representation for harmonic content in music signals. In: Int. Symp. on Music Information Retrieval (ISMIR), pp. 304–311 (2005)
6. Berenzweig, A., Logan, B., Ellis, D.P.W., Whitman, B.: A large scale evaluation of acoustic and subjective music similarity measures. In: Int. Symp. on Music Information Retrieval, ISMIR (2003)
7. Cano, P., Batlle, E., Kalker, T., Haitsma, J.: A review of audio fingerprinting. Journal of VLSI Signal Processing 41, 271–284 (2005)
8. Casey, M., Rhodes, C., Slaney, M.: Analysis of minimum distances in high-dimensional musical spaces. IEEE Trans. on Audio, Speech, and Language Processing 16(5), 1015–1028 (2008)
9. Casey, M., Slaney, M.: The importance of sequences in musical similarity. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), May 2006, vol. 5, p. V (2006)
10. Casey, M., Veltkamp, R.C., Goto, M., Leman, M., Rhodes, C., Slaney, M.: Content-based music information retrieval: current directions and future challenges. Proceedings of the IEEE 96(4), 668–696 (2008)
11. Dalla Bella, S., Peretz, I., Aronoff, N.: Time course of melody recognition: a gating paradigm study. Perception and Psychophysics 7(65), 1019–1028 (2003)

[20] http://www.pharos-audiovisual-search.eu

[21] http://www.variazioniproject.org

12. Dannenberg, R.B., Birmingham, W.P., Pardo, B., Hu, N., Meek, C., Tzanetakis, G.: A comparative evaluation of search techniques for query-by-humming using the musart testbed. Journal of the American Society for Information Science and Technology 58(5), 687–701 (2007)
13. de Cheveigné, A.: Pitch perception models. In: Plack, C.J., Oxenham, A., Fay, R.R., Popper, A.N. (eds.) Pitch – Neural coding and perception, pp. 169–233. Springer, New York (2005)
14. Deliège, I.: Cue abstraction as a component of categorisation processes in music listening. Psychology of Music 24(2), 131–156 (1996)
15. Dixon, S., Widmer, G.: Match: A music alignment tool chest. In: Int. Symp. on Music Information Retrieval (ISMIR), pp. 492–497 (2005)
16. Dowling, W.J.: Scale and contour: two components of a theory of memory for melodies. Psychological Review 85(4), 341–354 (1978)
17. Dowling, W.J., Harwood, J.L.: Music cognition. Academic Press, London (1985)
18. Downie, J.S.: The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research. Acoustical Science and Technology 29(4), 247–255 (2008)
19. Downie, J.S., Bay, M., Ehmann, A.F., Jones, M.C.: Audio cover song identification: MIREX 2006-2007 results and analyses. In: Int. Symp. on Music Information Retrieval (ISMIR), September 2008, pp. 468–473 (2008)
20. Egorov, A., Linetsky, G.: Cover song identification with IF-F0 pitch class profiles. In: MIREX extended abstract (September 2008)
21. Ellis, D.P.W., Cotton, C.: The 2007 labrosa cover song detection system. In: MIREX extended abstract (September 2007)
22. Ellis, D.P.W., Cotton, C., Mandel, M.: Cross-correlation of beat-synchronous representations for music similarity. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), April 2008, pp. 57–60 (2008)
23. Ellis, D.P.W., Poliner, G.E.: Identifying cover songs with chroma features and dynamic programming beat tracking. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), April 2007, vol. 4, pp. 1429–1432 (2007)
24. Ellis, D.P.W., Whitman, B., Berenzweig, A., Lawrence, S.: The quest for ground truth in musical artist similarity. In: Int. Symp. on Music Information Retrieval (ISMIR), October 2002, pp. 518–529 (2002)
25. Foote, J.: Arthur: Retrieving orchestral music by long-term structure. In: Int. Symp. on Music Information Retrieval (ISMIR) (October 2000)
26. Fujishima, T.: Realtime chord recognition of musical sound: a system using common lisp music. In: Int. Computer Music Conference (ICMC), pp. 464–467 (1999)
27. Gómez, E.: Tonal description of music audio signals. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain (2006), http://mtg.upf.edu/node/472
28. Gómez, E., Herrera, P.: The song remains the same: identifying versions of the same song using tonal descriptors. In: Int. Symp. on Music Information Retrieval (ISMIR), October 2006, pp. 180–185 (2006)
29. Gómez, E., Ong, B.S., Herrera, P.: Automatic tonal analysis from music summaries for version identification. In: Conv. of the Audio Engineering Society (AES) (October 2006); CD-ROM, paper no. 6902
30. Gouyon, F., Klapuri, A., Dixon, S., Alonso, M., Tzanetakis, G., Uhle, C., Cano, P.: An experimental comparison of audio tempo induction algorithms. IEEE Trans. on Speech and Audio Processing 14(5), 1832–1844 (2006)
31. Gusfield, D.: Algorithms on strings, trees and sequences: computer sciences and computational biology. Cambridge University Press, Cambridge (1997)

32. Harte, C.A., Sandler, M.B.: Automatic chord identification using a quantized chromagram. In: Conv. of the Audio Engineering Society (AES), pp. 28–31 (2005)
33. Heikkila, J.: A new class of shift-invariant operators. IEEE Signal Processing Magazine 11(6), 545–548 (2004)
34. Hu, N., Dannenberg, R.B., Tzanetakis, G.: Polyphonic audio matching and alignment for music retrieval. In: IEEE Workshop on Apps. of Signal Processing to Audio and Acoustics (WASPAA), pp. 185–188 (2003)
35. Izmirli, Ö.: Tonal similarity from audio using a template based attractor model. In: Int. Symp. on Music Information Retrieval (ISMIR), pp. 540–545 (2005)
36. Jensen, J.H., Christensen, M.G., Ellis, D.P.W., Jensen, S.H.: A tempo-insensitive distance measure for cover song identification based on chroma features. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), April 2008, pp. 2209–2212 (2008)
37. Jensen, J.H., Christensen, M.G., Jensen, S.H.: A chroma-based tempo-insensitive distance measure for cover song identification using the 2d autocorrelation. In: MIREX extended abstract (September 2008)
38. Kim, S., Narayanan, S.: Dynamic chroma feature vectors with applications to cover song identification. In: IEEE Workshop on Multimedia Signal Processing (MMSP), October 2008, pp. 984–987 (2008)
39. Kim, S., Unal, E., Narayanan, S.: Fingerprint extraction for classical music cover song identification. In: IEEE Int. Conf. on Multimedia and Expo (ICME), June 2008, pp. 1261–1264 (2008)
40. Kim, Y.E., Perelstein, D.: MIREX 2007: audio cover song detection using chroma features and hidden markov model. In: MIREX extended abstract (September 2007)
41. Kurth, F., Müller, M.: Efficient index-based audio matching. IEEE Trans. on Audio, Speech, and Language Processing 16(2), 382–395 (2008)
42. Larkin, C. (ed.): The Encyclopedia of Popular Music, 3rd edn. (November 1998)
43. Lee, K.: Identifying cover songs from audio using harmonic representation. In: MIREX extended abstract (September 2006)
44. Lee, K.: A system for acoustic chord transcription and key extraction from audio using hidden Markov models trained on synthesized audio. PhD thesis, Stanford University, USA (2008)
45. Lemstrom, K.: String matching techinques for music retrieval. PhD thesis, University of Helsinki, Finland (2000)
46. Levitin, D.: This is your brain on music: the science of a human obsession. Penguin (2007)
47. Manning, C.D., Prabhakar, R., Schutze, H.: An introduction to Information Retrieval. Cambridge University Press, Cambridge (2008), `http://www.informationretrieval.org`
48. Mardirossian, A., Chew, E.: Music summarization via key distributions: analyses of similarity assessment across variations. In: Int. Symp. on Music Information Retrieval, ISMIR (2006)
49. Marolt, M.: A mid-level melody-based representation for calculating audio similarity. In: Int. Symp. on Music Information Retrieval (ISMIR), October 2006, pp. 280–285 (2006)
50. Marolt, M.: A mid-level representation for melody-based retrieval in audio collections. IEEE Trans. on Multimedia 10(8), 1617–1625 (2008)
51. Miotto, R., Orio, N.: A music identification system based on chroma indexing and statistical modeling. In: Int. Symp. on Music Information Retrieval (ISMIR), September 2008, pp. 301–306 (2008)
52. Müller, M.: Information Retrieval for Music and Motion. Springer, Heidelberg (2007)

53. Müller, M., Kurth, F., Clausen, M.: Audio matching via chroma-based statistical features. In: Int. Symp. on Music Information Retrieval (ISMIR), pp. 288–295 (2005)
54. Myers, C.: A comparative study of several dynamic time warping algorithms for speech recognition. Master's thesis, Massachussets Institute of Technology, USA (1980)
55. Nagano, H., Kashino, K., Murase, H.: Fast music retrieval using polyphonic binary feature vectors. In: IEEE Int. Conf. on Multimedia and Expo (ICME), vol. 1, pp. 101–104 (2002)
56. Navarro, G., Mákinen, V., Ukkonen, E.: Algorithms for transposition invariant string matching. Journal of Algorithms (56) (2005)
57. Ong, B.S.: Structural analysis and segmentation of music signals. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain (2007), `http://mtg.upf.edu/node/508`
58. Oppenheim, A.V., Schafer, R.W., Buck, J.B.: Discrete-Time Signal Processing, 2nd edn. Prentice Hall, Englewood Cliffs (1999)
59. Pachet, F.: Knowledge management and musical metadata. Idea Group (2005)
60. Papadopoulos, H., Peeters, G.: Large-scale study of chord estimation algorithms based on chroma representation and hmm. In: Int. Conf. on Content-Based Multimedia Information, pp. 53–60 (2007)
61. Pickens, J.: Harmonic modeling for polyphonic music retrieval. PhD thesis, University of Massachussetts Amherst, USA (2004)
62. Poliner, G.E., Ellis, D.P.W., Ehmann, A., Gómez, E., Streich, S., Ong, B.S.: Melody transcription from music audio: approaches and evaluation. IEEE Trans. on Audio, Speech, and Language Processing 15, 1247–1256 (2007)
63. Purwins, H.: Proles of pitch classes. Circularity of relative pitch and key: experiments, models, computational music analysis, and perspectives. PhD thesis, Berlin University of Technology, Germany (2005)
64. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proc. of the IEEE (1989)
65. Rabiner, L.R., Juang, B.H.: Fundamentals of speech recognition. Prentice Hall, Englewood Cliffs (1993)
66. Riley, M., Heinen, E., Ghosh, J.: A text retrieval approach to content-based audio retrieval. In: Int. Symp. on Music Information Retrieval (ISMIR), September 2008, pp. 295–300 (2008)
67. Robine, M., Hanna, P., Ferraro, P., Allali, J.: Adaptation of string matching algorithms for identification of near-duplicate music documents. In: ACM SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN), pp. 37–43 (2007)
68. Sailer, C., Dressler, K.: Finding cover songs by melodic similarity. In: MIREX extended abstract (September 2006)
69. Sankoff, D., Kruskal, J.: Time warps, string edits, and macromolecules. Addison-Wesley, Reading (1983)
70. Scaringella, N., Zoia, G., Mlynek, D.: Automatic genre classification of music content: a survey. IEEE Signal Processing Magazine 23(2), 133–141 (2006)
71. Schellenberg, E.G., Iverson, P., McKinnon, M.C.: Name that tune: identifying familiar recordings from brief excerpts. Psychonomic Bulletin and Review 6(4), 641–646 (1999)
72. Schulkind, M.D., Posner, R.J., Rubin, D.C.: Musical features that facilitate melody identification: how do you know it's your song when they finally play it? Music Perception 21(2), 217–249 (2003)
73. Selfridge-Field, E.: Conceptual and representational issues in melodic comparison. MIT Press, Cambridge (1998)

74. Serrà, J., Serra, X., Andrzejak, R.G.: Cross recurrence quantification for cover song identification. New Journal of Physics 11, art. 093017 (September 2009)
75. Serrà, J., Gómez, E., Herrera, P.: Transposing chroma representations to a common key. In: IEEE CS Conference on The Use of Symbols to Represent Music and Multimedia Objects, October 2008, pp. 45–48 (2008)
76. Serrà, J., Gómez, E., Herrera, P., Serra, X.: Chroma binary similarity and local alignment applied to cover song identification. IEEE Trans. on Audio, Speech, and Language Processing 16(6), 1138–1152 (2008)
77. Sheh, A., Ellis, D.P.W.: Chord segmentation and recognition using em-trained hidden markov models. Int. Symp. on Music Information Retrieval (ISMIR), pp. 183–189 (2003)
78. Tsai, W.H., Yu, H.M., Wang, H.M.: A query-by-example technique for retrieving cover versions of popular songs with similar melodies. In: Int. Symp. on Music Information Retrieval (ISMIR), pp. 183–190 (2005)
79. Tsai, W.H., Yu, H.M., Wang, H.M.: Using the similarity of main melodies to identify cover versions of popular songs for music document retrieval. Journal of Information Science and Engineering 24(6), 1669–1687 (2008)
80. Tversky, A.: Features of similarity. Psychological Review 84, 327–352 (1977)
81. Typke, R.: Music retrieval based on melodic similarity. PhD thesis, Utrecht University, Netherlands (2007)
82. Tzanetakis, G.: Pitch histograms in audio and symbolic music information retrieval. In: Int. Symp. on Music Information Retrieval (ISMIR), pp. 31–38 (2002)
83. Ukkonen, E., Lemstrom, K., Mäkinen, V.: Sweepline the music! Comp. Sci. in Perspective, 330–342 (2003)
84. Unal, E., Chew, E.: Statistical modeling and retrieval of polyphonic music. In: IEEE Workshop on Multimedia Signal Processing (MMSP), pp. 405-409 (2007)
85. Vignoli, F., Paws, S.: A music retrieval system based on user-driven similarity and its evaluation. In: Int. Symp. on Music Information Retrieval (ISMIR), pp. 272–279 (2005)
86. Voorhees, E.M., Harman, D.K.: Trec: Experiment and evaluation in information retrieval (2005)
87. White, B.W.: Recognition of distorted melodies. American Journal of Psychology 73, 100–107 (1960)
88. Xu, R., Wunsch, D.C.: Clustering. IEEE Press, Los Alamitos (2009)
89. Yang, C.: Music database retrieval based on spectral similarity. Technical report (2001)
90. Yu, Y., Downie, J.S., Chen, L., Oria, V., Joe, K.: Searching musical audio datasets by a batch of multi-variant tracks. In: ACM Multimedia, October 2008, pp. 121–127 (2008)
91. Yu, Y., Downie, J.S., Mörchen, F., Chen, L., Joe, K., Oria, V.: Cosin: content-based retrieval system for cover songs. In: ACM Multimedia, October 2008, pp. 987–988 (2008)