OPEN ACCESS

Cross recurrence quantification for cover song identification

New Journal of Physics
The open-access journal for physics

Cross recurrence quantification for cover song identification

Joan Serrà[1], **Xavier Serra and Ralph G Andrzejak**
Department of Information and Communication Technologies,
Universitat Pompeu Fabra, Roc Boronat 138, 08018 Barcelona, Spain
E-mail: joan.serraj@upf.edu

New Journal of Physics **11** (2009) 093017 (20pp)
Received 22 July 2009
Published 15 September 2009
Online at http://www.njp.org/
doi:10.1088/1367-2630/11/9/093017

Abstract. There is growing evidence that nonlinear time series analysis techniques can be used to successfully characterize, classify, or process signals derived from real-world dynamics even though these are not necessarily deterministic and stationary. In the present study, we proceed in this direction by addressing an important problem our modern society is facing, the automatic classification of digital information. In particular, we address the automatic identification of cover songs, i.e. alternative renditions of a previously recorded musical piece. For this purpose, we here propose a recurrence quantification analysis measure that allows the tracking of potentially curved and disrupted traces in cross recurrence plots (CRPs). We apply this measure to CRPs constructed from the state space representation of musical descriptor time series extracted from the raw audio signal. We show that our method identifies cover songs with a higher accuracy as compared to previously published techniques. Beyond the particular application proposed here, we discuss how our approach can be useful for the characterization of a variety of signals from different scientific disciplines. We study coupled Rössler dynamics with stochastically modulated mean frequencies as one concrete example to illustrate this point.

[1] Author to whom any correspondence should be addressed.

Contents

1. Introduction

An unprecedented growth in the availability of and access to digital information is taking place in today's society, and music is a paradigmatic example. Online digital music collections are in the order of millions of tracks, and personal collections can easily exceed the practical limits on the time to listen to them [1]. This huge amount of information readily accessible for end users poses major challenges for automatically describing, understanding, searching, retrieving, and organizing musical contents. Music information retrieval (MIR) is the interdisciplinary research field that deals with these challenges [2].

MIR systems use multiple sources of information: the raw audio signal, symbolic music representations, audio metadata, tags provided by users or experts, music and social networks data, etc. In content-based MIR, much effort is focused on extracting information from the raw audio signal to represent certain musical aspects such as timbre, melody, main tonality, chords, or tempo [1]. Usually, these features are computed in a short-time moving window either from a temporal, spectral, or cepstral representation of the audio signal [1], leading to a descriptor time series reflecting the temporal evolution of a given musical aspect. While common MIR strategies characterize these time series by means of statistical modeling or machine learning techniques [3]–[5], raw descriptor time series are used for many tasks such as audio alignment and matching [6], song structure analysis [7], music similarity [8], audio fingerprinting [9], or cover song identification [10]–[18].

A cover song is an alternative version, performance, rendition, or recording of a previously recorded musical piece. While cover songs might differ from their originals in several musical aspects such as timbre, tempo, song structure, main tonality, arrangement, lyrics, or language of the vocals, they resemble their originals with regard to other features. A robust so-called 'mid-level feature' that is largely preserved under the mentioned musical variations is the

tonal sequence. Tonal sequences can be understood as series of different notes. These notes can be played alone for each time slot (a melody) or can be played simultaneously with other notes (chord or harmonic progressions). Methods for automatic cover song identification usually exploit tonal sequence similarity and attempt to be robust against common changes in other musical aspects [18]. In general, they either aim to extract the predominant melody, a chord progression, or a chroma time series (a mid-level feature representing harmonic content) from the raw audio signal and make it independent of the main tonality. Then, for obtaining a similarity measure between songs, tonality descriptor time series are usually compared by means of techniques like dynamic time warping, edit-distance variants, string matching algorithms, subsequence hashing, or by common similarity functions (for an overview see [18]).

Cover song identification has recently become a very active area of study in the MIR community [10]–[18]. From a research point of view, cover song identification is a task where the relation between songs is context-independent and can be quantitatively defined and objectively measured. It expands the notions of music similarity beyond acoustic resemblance to include the important idea that musical works retain their identity despite variations in many musical aspects [20]. From a practical and commercial point of view, quantifying music similarity is the key to automatically searching and organizing music collections. Furthermore, identifying cover songs has a direct implication to musical rights management and licenses. In addition, from a user's point of view, finding all versions of a particular song can be valuable and fun.

The MIR evaluation exchange (MIREX) is an international community-based framework for the formal evaluation of MIR systems and algorithms [21]. Among other tasks, MIREX allows comparing different algorithms for artist identification, genre classification, or music transcription[2]. In particular, MIREX allows for an objective assessment of the accuracy of different cover song identification algorithms. For that purpose, participants can submit their algorithms as binary executables, and the MIREX organizers determine and publish the algorithms' accuracies and runtimes. The underlying music collections are never published or disclosed to the participants, either before or after the contest. Therefore, participants cannot tune their algorithms to the music collections used in the evaluation process.

For the 2007 edition of the MIREX cover song identification contest, our group submitted an algorithm that we subsequently described in [17]. This algorithm, which used a specifically designed chroma similarity measure and a subsequence matching method, yielded the highest accuracy of all algorithms submitted in 2007 and in earlier editions. For the 2008 edition, we used a qualitatively novel approach. The cover song identification measure that we derived from this approach ($Q_{\max}$) and a composition of this measure with a simple post-processing step ($Q^*_{\max}$) yielded the two highest accuracies of all algorithms submitted in 2008 and in earlier editions. In particular, the accuracy of both $Q_{\max}$ and $Q^*_{\max}$ clearly surpassed our earlier algorithm proposed in [17].

The $Q_{\max}$ algorithm was submitted to the MIREX contest as a binary executable, and we here disclose for the first time the underlying procedure. While this algorithm shares MIR pre-processing steps with [17], the crucial difference is that it involves techniques derived from nonlinear time series analysis [22]. More specifically, $Q_{\max}$ is a recurrence quantification analysis (RQA) measure [23]–[26] that is extracted from cross recurrence plots (CRPs) [27], which are the bivariate generalization of classical recurrence plots (RPs) [28]. The framework

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

of nonlinear time series analysis offers a variety of techniques to quantify similarities between dynamics based on signals measured from them. Among these techniques, the CRP seems most suitable to analyze pairs of musical descriptor time series since it is defined for pairs of signals of different lengths and can easily cope with variations in the timescale and non-stationarities of the dynamics [29, 30]. Here, we construct CRPs from delay coordinate state space representations of multivariate descriptor time series of songs.

CRPs and RQA measures are known as very intuitive and powerful tools in various disciplines such as astrophysics, earth sciences, engineering, biology, cardiology, or neuroscience (see [26] and references therein). However, to the best of our knowledge, there are no previous applications of CRPs and RQA measures to musical signals. In general, only few studies apply nonlinear time series analysis to musical signals. In [31, 32], delay coordinates are applied to raw audio signals with regard to audio analysis and visualization. In [33]–[35], delay coordinates are applied to musical descriptor time series with regard to genre classification, user preferences, and timbre modeling. In [36], delay coordinates are applied to human speech signals for the purpose of local projective noise reduction. Subsequently, in [37], an RQA measure was defined to automatically adjust the best neighborhood size for this local projection.

It should be noted that RPs and CRPs have certain analogies with commonly used MIR methods. In particular, the so-called self-similarity matrix was introduced in [38] to visualize music and audio tracks and later used in [39] for song structure segmentation or in [40] for identifying components of an audio piece. Currently, self-similarity matrices are commonly used for diverse tasks such as song structure analysis [7] or musical meter detection [41]. Cross similarity matrices are used, either directly or indirectly, in audio matching algorithms [6] and in some cover song identification methods [18]. However, in contrast to CRPs, these similarity matrices do not apply any delay coordinate state space representation and are, in general, not thresholded.

A brief overview of the $Q_{\max}$ algorithm and the resulting structure of this paper can be outlined as follows. Given two songs, we first extract their chroma descriptor time series and transpose one song to the main tonality of the other (section 2.1). From this pair of multivariate time series, we form state space representations of the two songs using delay coordinates involving an embedding dimension m and time delay τ (section 2.2). From this state space representation, we construct a CRP using a fixed maximum percentage of nearest neighbors κ (section 2.3). Subsequently, we use $Q_{\max}$ to extract features that are sensitive to cover song CRP characteristics, which results in two additional parameters γ_{o} and γ_{e}. In particular, we derive $Q_{\max}$ from a previously published RQA measure ($L_{\max}$, [28]), but adapt it in two steps (via $S_{\max}$) to the problem at hand (section 2.4). We evaluate our approach using a large collection of musical pieces (section 3.1). This music collection was compiled prior to and independently from the present study and our participation in the MIREX contest. We use a subset of this music collection and a standard information retrieval (IR) evaluation methodology (section 3.2) to, at first, perform an in-sample optimization of parameters m, τ, κ, γ_{o} and γ_{e} (section 4.1). We subsequently report the out-of-sample accuracy with optimized parameters of $L_{\max}$, $S_{\max}$, and $Q_{\max}$ in identifying cover songs (section 4.2). All these steps were carried out before we submitted the resulting algorithm to the 2008 MIREX cover song identification contest as a further out-of-sample validation. We review results of this 2008 and the 2007 editions (section 4.3) before we draw our conclusions (section 5). As an outlook (section 6), we provide concrete perspectives for future applications of our technique. For this purpose we use coupled Rössler dynamics with stochastically modulated mean frequencies.

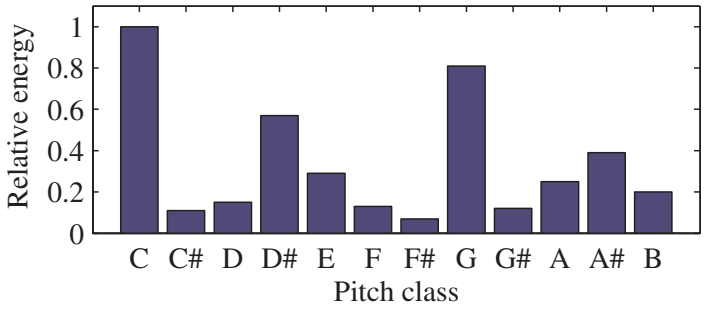

Figure 1. Example of a PCP feature vector extracted from an audio window of 464 ms. This PCP corresponds to a C minor chord environment (it mostly contains C, D# and G pitch classes), where the root pitch class (C) is predominant.

2. Method

2.1. Pre-processing

The tonal sequence is the most important characteristic shared among covers. To estimate tonal sequences of musical pieces one can employ chroma or pitch class profile (PCP) features. These are widely used in the MIR community [42]–[45] and are proven to work well as primary information for cover song identification systems [18]. For systems employing PCP see [10, 13, 15, 16, 17, 19].

In general, PCP features are robust against non-tonal components (e.g. ambient noise or percussive sounds) and independent of timbre and the specific instruments used [45]. Furthermore, they are independent of a musical piece's loudness and volume fluctuations. PCP features are derived from the frequency dependent energy in a given range (typically from 50 to 5000 Hz) in short-time spectral representations (e.g. 100 ms) of audio signals computed in a moving window. This energy is usually mapped into an octave-independent histogram representing the relative intensity of each of the 12 semitones of the western music chromatic scale (12 pitch classes). To normalize with respect to loudness, this histogram can be divided by its maximum value, thus leading to values between 0 and 1 (figure 1).

We here use harmonic PCPs (HPCPs) [45]. These features share the aforementioned PCP properties, but are based only on the peaks of the spectrum within a certain frequency band, thereby they reduce the influence of noisy spectral components. Furthermore, HPCPs are tuning independent, so that the reference tone can be different from the standard tone A 440 Hz. In addition, they take into account the presence of harmonic frequencies. Except for that we here use 12 instead of 36 HPCP bins, we use the same HPCP extraction procedure and parameters as in [17], to which we refer for further details.

The computation of HPCPs in a moving window results in a multidimensional time series x for each song, expressing its temporal tonal evolution $x = \{x_{h,i}\}$ for $h = 1, \ldots, H$ and $i = 1, \ldots, N_x^*$, where $H = 12$ is a routinely employed number of HPCP bins [42]–[45] and N_x^* represents the total number of windows (figure 2). We here use windows of 464 ms with no overlap between subsequent windows.

The last pre-processing step consists in transposing one HPCP time series to the main tonality of the other. A change in the main tonality is a common alteration when musicians

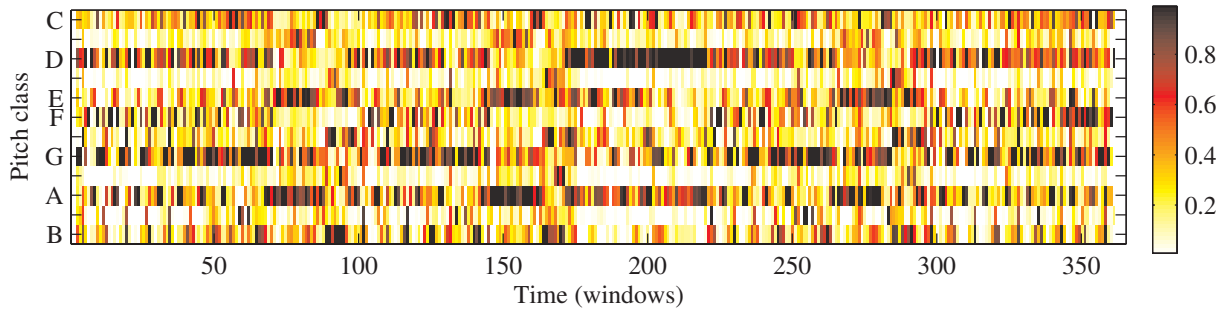

Figure 2. Example of an HPCP time series extracted using a moving window from the song *Day Tripper* as performed by The Beatles.

perform cover versions. This is usually done to adapt the original composition to a different singer or solo instrument, or just for aesthetic reasons. In HPCP representations, a change in the main tonality is represented by a circular pitch class shift. Accordingly, one can reverse this change using an appropriate circular shift of the pitch class components along the vertical axis of an HPCP time series (e.g. to transpose the time series depicted in figure 2 from D to C, one has to shift the pitch class components circularly up by two bins, i.e. two semitones, for all windows). To determine the number of bins to transpose, we use the optimal transposition index procedure proposed in [17] and extended and further evaluated in [46].

2.2. State space embedding

An HPCP time series is a multivariate representation of the temporal tonal evolution of a given song X. Certainly, it does not represent a signal measured from a stationary dynamical system which could be described by some equation of motion. Nonetheless, delay coordinates [47], a tool that is routinely used in nonlinear time series analysis [22], can be pragmatically employed to facilitate the extraction of information contained in an HPCP time series x (cf [36, 37]). In particular, by evaluating vectors of sample sequences, delay coordinates allow one to assess systems recurrences more reliably than by using only the scalar samples. One should note that such a use of sequences of notes instead of isolated ones is essential in music [48] and is important for melody perception and recognition [49].

Considering the temporal evolution of each individual pitch class, we construct a time series of delay coordinate state space vectors $\mathbf{x} = \{\mathbf{x}_i\}$ for $i = 1, \ldots, N_x$, with $N_x = N_x^* - (m-1)\tau$ and

$$\mathbf{x}_i = (x_{1,i}, x_{1,i+\tau}, \ldots x_{1,i+(m-1)\tau}, x_{2,i}, x_{2,i+\tau}, \ldots x_{2,i+(m-1)\tau}, \ldots x_{H,i}, x_{H,i+\tau}, \ldots x_{H,i+(m-1)\tau}), \qquad (1)$$

where m is the unitless embedding dimension, and τ is the time delay in units of the number of windows. For nonlinear time series analysis, an appropriate choice of m and τ is crucial to extract meaningful information from noisy signals of finite length [22]. While recipes for the estimation of optimal fixed values of m and τ exist (e.g. the false nearest neighbors' method and the use of the auto-correlation function decay time [22]), we here study cover song identification accuracy under variation of these parameters and select the best combination (section 4).

2.3. CRP

An RP is a straightforward way to visualize characteristics of similar system states attained at different times [28]. For this purpose, two discrete time axes span a square matrix which is filled with zeros and ones, typically visualized as white and black cells, respectively. Each black cell at coordinates (i, j) indicates a recurrence, i.e. a state at time i which is similar to a state at time j. Thereby, the main diagonal line is black. CRPs are constructed in the same way as RPs, but now the two axes span a rectangular, not necessarily square matrix [27]. A CRP allows one to highlight equivalences of states between two systems attained at different times. When a CRP is used to characterize distinct systems, the main diagonal is, in general, not black, and any diagonal path of connected black cells represents similar state sequences exhibited by both systems [26].

To analyze dependencies between two different signals x and y, here representing two songs, we compute a CRP R from

$$R_{i,j} = \Theta(\varepsilon_i^x - \|\mathbf{x}_i - \mathbf{y}_j\|)\Theta(\varepsilon_j^y - \|\mathbf{x}_i - \mathbf{y}_j\|) \tag{2}$$

for $i = 1, \ldots, N_x$ and $j = 1, \ldots, N_y$, where $\mathbf{x}_i$ and $\mathbf{y}_j$ are state space representations of songs X and Y at windows i and j, respectively, $\Theta(\cdot)$ is the Heaviside step function ($\Theta(v) = 0$ if $v < 0$ and $\Theta(v) = 1$ otherwise), ε_i^x and ε_j^y are two different threshold distances, and $\|\cdot\|$ is some norm. We here use the Euclidean norm. Note that by equation (2) $R_{i,j} = 1$ if and only if $\mathbf{x}_i$ is a neighbor of $\mathbf{y}_j$ and $\mathbf{y}_j$ is a neighbor of $\mathbf{x}_i$.

The thresholds ε_i^x and ε_j^y are adjusted such that a maximum percentage of neighbors κ is used for both $\mathbf{x}_i$ and $\mathbf{y}_j$. In this way, the total number of nonzero entries in each row and column never exceeds κN_y and κN_x, respectively. In-line with studies on the identification of deterministic signals in noisy environments [27], in pre-analysis we found the use of a fixed percentage of neighbors κ superior to the use of a fixed threshold ε. We study the influence of the parameter κ in section 4.

In general, pairs of unrelated songs result in CRPs that exhibit no evident structure, while CRPs constructed for two cover songs show distinct extended patterns (figure 3). These extended patterns usually correspond to similar sections, phrases, or progressions between both musical pieces X and Y.

2.4. Recurrence quantification measures for cover song identification

Given a CRP representation of two songs, we require a quantitative criterion to determine whether they are covers or not. In pre-analysis, we tested different RQA measures [26] as input for binary classifiers such as trees or support vector machines in combination with several feature selection algorithms[3] [50]. This analysis showed that the maximal length of diagonal lines ($L_{\max}$) feature yielded by far the highest discriminative power between CRPs from covers and non-covers. All other RQA measures that we tried (recurrence rate, determinism, average diagonal length, entropy, ratio, laminarity, trapping time, maximal length of horizontal or vertical lines [26], and combinations of them) were found to have no or very low discriminative power.

[3] We use the Weka data mining software: http://www.cs.waikato.ac.nz/ml/weka

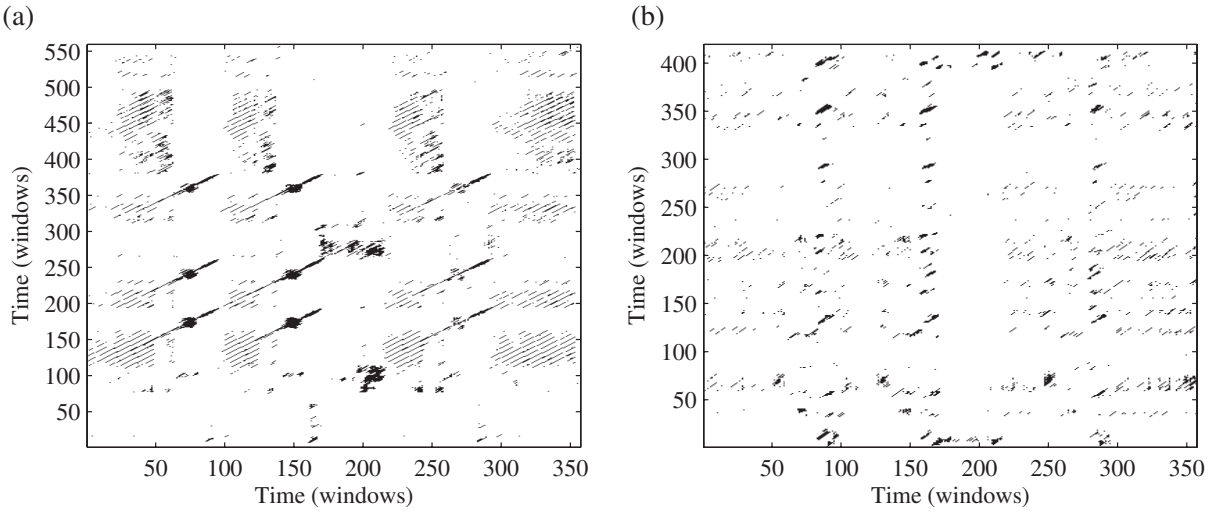

Figure 3. CRPs for the song *Day Tripper* as performed by The Beatles, taken as song X, versus two different songs, taken as song Y. These are a cover made by the group Ocean Colour Scene (a) and the song *I've Got a Crush on You* as performed by Frank Sinatra (b). Parameters are $m = 9$, $\tau = 1$, and $\kappa = 0.08$.

The $L_{\max}$ measure introduced in [28] can be expressed as the maximum value of a cumulative matrix L computed from the CRP. We initialize $L_{1,j} = L_{i,1} = 0$ for $i = 1, \ldots, N_x$ and $j = 1, \ldots, N_y$, and then recursively apply

$$L_{i,j} = \begin{cases} L_{i-1,j-1} + 1, & \text{if } R_{i,j} = 1, \\ 0, & \text{if } R_{i,j} = 0, \end{cases} \tag{3}$$

for $i = 2, \ldots, N_x$ and $j = 2, \ldots, N_y$, and define $L_{\max} = \max\{L_{i,j}\}$ for $i = 1, \ldots, N_x$ and $j = 1, \ldots, N_y$.

To understand why $L_{\max}$ is performing so well we depict some example CRPs, where we use the same song for X and three different songs for Y (figure 4). A high $L_{\max}$ value is obtained when X and Y are covers (figure 4(a)), whereas a low value is obtained when that is not the case (figure 4(c)). An intermediate value is obtained for two songs that share a common tonal progression, but only for brief periods (figure 4(b)). It turns out that this particular example of figure 4(b) is a border case where one would consider the two songs to be covers or not. The two songs are very different even in terms of main melody and tonality, but still they share a very characteristic sample featuring a flute hook that forms the basis of both songs[4].

Diagonal patterns are clearly discernible in figures 4(a) and (b), and the longest of these diagonals corresponds to the maximum time that X and Y evolve together without disruptions (i.e. the maximal length of their shared tonal sequence). Note that only in figure 4(a) the longest diagonal is found close to the main diagonal. However, that is not a necessary criterion of Y being a cover of X (figure 4(b)). In general, this depends on the musical structure of the cover song. Often, new performers add, delete, or change the introduction,

[4] http://news.bbc.co.uk/2/hi/entertainment/4354028.stm

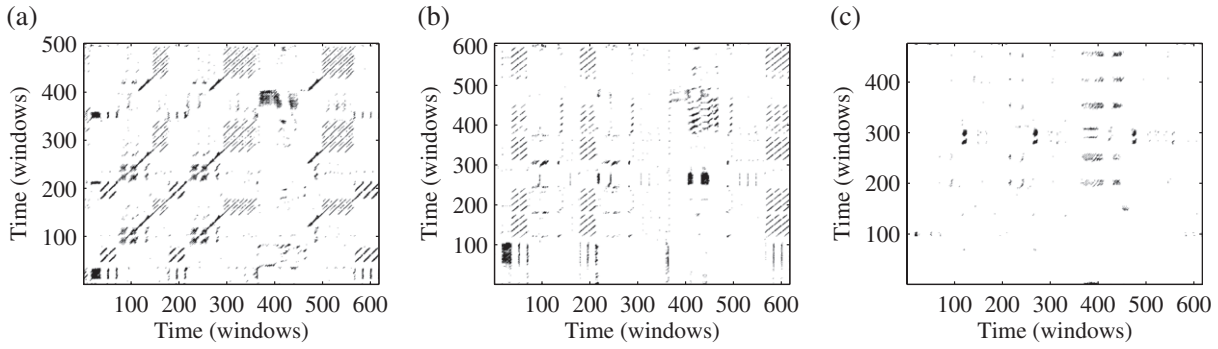

Figure 4. CRPs for the song *Gimme, Gimme, Gimme* as performed by the group ABBA, taken as song X, versus three different songs, taken as song Y. These are a cover made by the group A-Teens (a), a techno performance of the song *Hung up* by Madonna (b), and the song *The Robots* by Kraftwerk (c). In (a) $L_{\max} = 43$ starting at windows (118, 121), in (b) $L_{\max} = 34$ starting at windows (176, 130), and in (c) $L_{\max} = 16$ starting at windows (373, 245). Parameters are the same as in figure 3.

solo sections, endings, verses, and so forth. Thus, to account for structure changes, it is necessary to consider any diagonal regardless of its position in the CRP. This allows one to detect passages of a song that have been inserted in any part of another song. However, while $L_{\max}$ can account for such structural changes, it cannot account for tempo changes. When covering a musical piece, musicians often adapt the tempo to their needs and, even in a live performance of the original artist, this feature can change with respect to the original recording. Tempo deviations between two cover songs result in the curving of CRP diagonal traces.

To quantify the length of curved traces we therefore extend equation (3) and compute a cumulative matrix S from the CRP. We initialize $S_{1,j} = S_{2,j} = S_{i,1} = S_{i,2} = 0$ for $i = 1, \ldots, N_x$ and $j = 1, \ldots, N_y$, and then recursively apply

$$S_{i,j} = \begin{cases} \max\{S_{i-1,j-1}, S_{i-2,j-1}, S_{i-1,j-2}\} + 1, & \text{if } R_{i,j} = 1, \\ 0, & \text{if } R_{i,j} = 0, \end{cases} \tag{4}$$

for $i = 3, \ldots, N_x$ and $j = 3, \ldots, N_y$. Here, the maximum value $S_{\max} = \max\{S_{i,j}\}$ for $i = 1, \ldots, N_x$ and $j = 1, \ldots, N_y$, corresponds to the length of the longest curved trace in the CRP. This formulation is inspired by common alignment algorithms [51, 52], but constrains the possible alignments by excluding horizontal and vertical paths. We should note that these particular path connections ($S_{i-1,j-1}$, $S_{i-2,j-1}$, $S_{i-1,j-2}$), which are only one aspect of equation (4), were used before. They were found to work well for speech recognition in application to distance matrices [53], and for cover song identification in application to the so-called optimal transposition index-based binary similarity matrices [17].

Apart from tempo deviations, musicians might skip some chords or part of the melody when performing cover songs. This practice leads to short disruptions in otherwise coherent traces (see, e.g. figure 3(a)). Moreover, such disruptions can also be caused by the fact that HPCP features might contain some energy not directly associated to tonal content. To account for disruptions, we therefore extend equation (4) and compute a cumulative matrix Q from

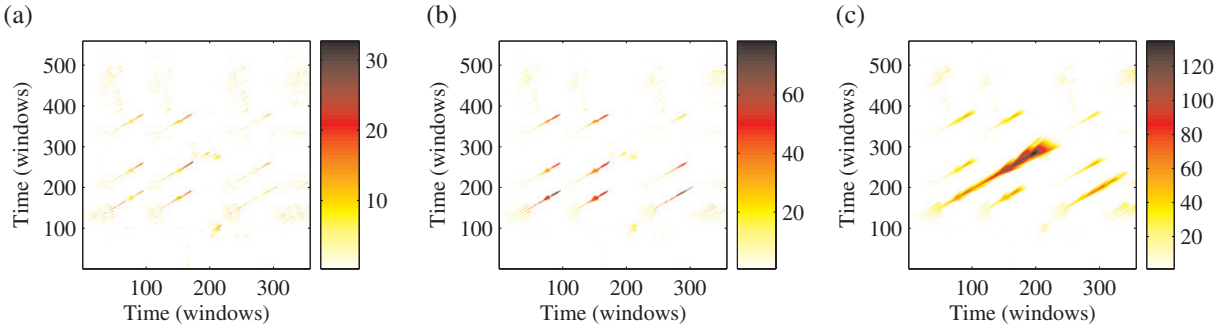

Figure 5. *Day Tripper* as performed by The Beatles, taken as song X, versus Ocean Colour Scene performance, taken as song Y. Example plots of L (a), S (b) and Q (c). Note the increase in the maximum values (colorscales). In (a) $L_{\max} = 33$ starting at windows (140, 232), in (b) $S_{\max} = 79$ starting at windows (216,142), and in (c) $Q_{\max} = 136$ starting at windows (14, 118). CRP parameters are the same as in figure 3. Parameters for (c) are $\gamma_o = 3$ and $\gamma_e = 7$.

the CRP. We initialize $Q_{1,j} = Q_{2,j} = Q_{i,1} = Q_{i,2} = 0$ for $i = 1, \ldots, N_x$ and $j = 1, \ldots, N_y$, and then recursively apply

$$Q_{i,j} = \begin{cases} \max\{Q_{i-1,j-1}, Q_{i-2,j-1}, Q_{i-1,j-2}\} + 1, & \text{if } R_{i,j} = 1, \\ \max\{0, Q_{i-1,j-1} - \gamma(R_{i-1,j-1}), & \\ \quad Q_{i-2,j-1} - \gamma(R_{i-2,j-1}), & \\ \quad Q_{i-1,j-2} - \gamma(R_{i-1,j-2})\}, & \text{if } R_{i,j} = 0, \end{cases} \tag{5}$$

for $i = 3, \ldots, N_x$ and $j = 3, \ldots, N_y$, with

$$\gamma(z) = \begin{cases} \gamma_o, & \text{if } z = 1, \\ \gamma_e & \text{if } z = 0. \end{cases} \tag{6}$$

Hence γ_o is a penalty for a disruption onset and γ_e is a penalty for a disruption extension. The zero inside the second max clause in equation (5) is used to prevent that these penalties lead to negative entries of Q. Note that for $\gamma_o, \gamma_e \to \infty$, equation (5) becomes equation (4). For $\gamma_o = \gamma_e = 0$, $Q_{i,j}$ becomes a cumulative value indicating global similarity between two time series starting at sample 0 and ending at samples i and j, respectively. Note that this has certain analogies with classical dynamic time warping algorithms [51]. Instead of fixing γ_o and γ_e *a priori*, we study their influence on the accuracy of our cover song identification system (section 4). Analogously to $L_{\max}$ and $S_{\max}$, we take $Q_{\max} = \max\{Q_{i,j}\}$ for $i = 1, \ldots, N_x$ and $j = 1, \ldots, N_y$ to quantify the length of the longest curved and potentially disrupted trace in the CRP.

For illustration we depict some examples for the three quantification measures discussed in this section (figure 5). The $L_{\max}$ measure (figure 5(a)) characterizes straight diagonals regardless of their position. The $S_{\max}$ measure can account for tempo fluctuations resulting in curved traces (figure 5(b)). Furthermore, the $Q_{\max}$ measure allows for disruptions of the tonal progression (figure 5(c)).

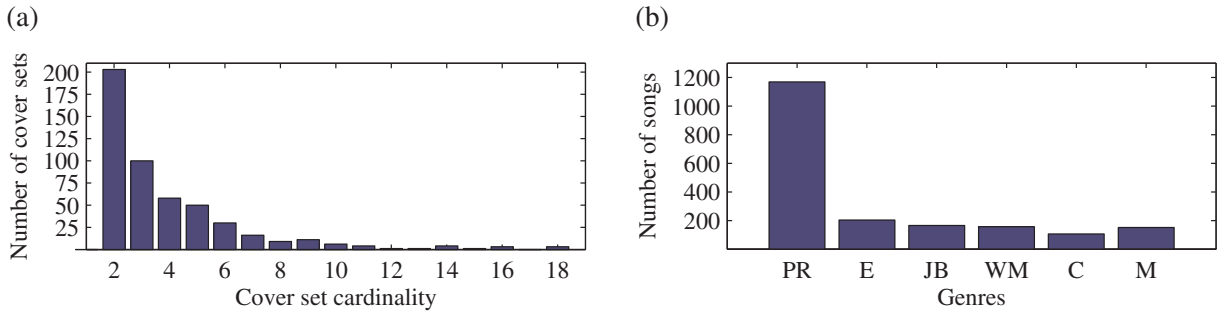

Figure 6. Distribution of the cover set cardinality (a) and the distribution of genres across all songs (b). PR stands for pop-rock, E for electronic, JB for jazz-blues, WM for world music, C for classical and M for miscellaneous.

3. Evaluation

3.1. Evaluation data

To test the effectiveness of the implemented approaches, we analyze a music collection comprising a total of 1953 commercial songs with an average song length of 3.5 min, ranging from 0.5 to 7 min. These songs include 500 cover sets, where cover set refers to a group of versions of the same song. The average cardinality of these cover sets (i.e. the number of songs per cover set) is 3.9, ranging from 2 to 18 (figure 6(a)). In composing this music collection we aimed at including a variety of styles and genres (figure 6(b)). No further criterion for the inclusion or exclusion of songs was applied. A complete list of the music collection can be found (http://mtg.upf.edu/people/jserra/). This music collection was compiled prior to and independently from the present study.

In order to form a training and three testing music collections, we split the total number of 500 cover sets into three non-overlapping subsets. The training collection contains 90 songs consisting of 15 cover sets of cardinality 6. The first testing collection contains 330 songs divided into 30 cover sets of cardinality 11. The second testing collection contains the remaining 455 cover sets each having cardinalities between 2 and 18, resulting in a total of 1533 songs. A further testing collection is defined as the union of first and second testing collections.

3.2. Evaluation methodology

Given a music collection with D songs, we calculate $L_{\max}$, $S_{\max}$ and $Q_{\max}$ for all $\frac{D(D-1)}{2}$ possible pairwise combinations. Once such a similarity matrix is computed as primary source of information, we can resort to standard IR measures to evaluate the discriminative power of this information. We use the mean of average precision measure [54], which we denote as Ψ. To calculate this measure, the similarity matrix is used to compute a list Λ_q of $D-1$ songs sorted in descending order with regard to their similarity to song q. Suppose that the query song q belongs to a cover set comprising $C_q + 1$ songs. Then, the average precision ψ_q is obtained as

$$\psi_q = \frac{1}{C_q} \sum_{r=1}^{D-1} P_q(r) I_q(r), \tag{7}$$

where $P_q(r)$ is the precision of the sorted list Λ_q at rank r,

$$P_q(r) = \frac{1}{r}\sum_{l=1}^{r} I_q(l), \tag{8}$$

and $I_q(\cdot)$ is the so-called relevance function ($I_q(u) = 1$ if the song with rank u in the sorted list is a cover of q, and $I_q(u) = 0$ otherwise). Hence ψ_q ranges between 0 and 1. If the cover songs take the first C_q ranks, we obtain $\psi_q = 1$. If all cover songs are found towards the end of Λ_q, we obtain values close to 0. The Ψ measure is calculated as the mean of average precisions ψ_q across all queries q. This evaluation measure is routinely employed in a wide variety of tasks in the IR [54] and MIR communities, including the MIREX cover song identification task [20]. Using equations (7) and (8) has the advantage of taking into account the whole sorted list where correct items with low rank receive the largest weights.

Additionally, we estimate the accuracy level expected under the null hypothesis that the similarity matrix has no discriminative power with regard to the assignment of cover sets. For this purpose, we separately permute Λ_q for all q and all other steps remain the same. We repeat this process 19 times, corresponding to a significance level of 0.05 of this Monte Carlo null hypothesis test, and take the average, resulting in Ψ_{null}. This Ψ_{null} can be used to estimate the accuracy of all measures $L_{\max}$, $S_{\max}$ and $Q_{\max}$ under the specified null hypothesis.

4. Results

4.1. Parameter optimization

We use the training collection to study the influence of the embedding parameters m and τ and the percentage of nearest neighbors κ on our accuracy measure Ψ. Results for $Q_{\max}$ (figure 7) illustrate that the use of an embedding ($m > 1$) improves the accuracy of the algorithm as compared to no embedding ($m = 1$). A broad peak of near-maximal Ψ values is established for a considerable range of embedding windows (approximately $7 < (m-1)\tau < 17$). From these near-maximal values, Ψ decreases weakly upon further increasing of the embedding window. Optimal κ values are found between 0.05 and 0.15. Therefore, within these broad ranges of the embedding window $(m-1)\tau$ and κ values, no fine tuning of any of the parameters is required to yield near-optimal accuracy. In the following we use $m = 10$, $\tau = 1$ and $\kappa = 0.1$.

While accuracies shown in figure 7 are computed for a disruption onset $\gamma_o = 2$ and disruption extension $\gamma_e = 2$ penalties, the influence of these penalty parameters is further studied in figure 8. Recall that γ_o and γ_e are introduced only in the definition of $Q_{\max}$ and that for $\gamma_o, \gamma_e \to \infty$, the measure $Q_{\max}$ (equation (5)) reduces to $S_{\max}$ (equation (4)). Using finite values of these terms generally increases the accuracy, revealing the advantage of $Q_{\max}$ over $S_{\max}$. Optimal $Q_{\max}$ accuracy values are found for $\gamma_o = 5$ and $\gamma_e = 0.5$.

The same parameter optimization described above for $Q_{\max}$ was carried out separately for $L_{\max}$ and $S_{\max}$, and $m = 10$, $\tau = 1$ and $\kappa = 0.1$ led to near-optimal accuracies also for these measures. Furthermore, no fine tuning was required since iso-τ and iso-m curves for different κ values have similar shapes as the ones depicted for $Q_{\max}$ in figure 7. For the training collection, this in-sample parameter optimization leads to the following accuracies (figure 9(a)): $\Psi_{L_{\max}} = 0.640$, $\Psi_{S_{\max}} = 0.728$ and $\Psi_{Q_{\max}} = 0.813$.

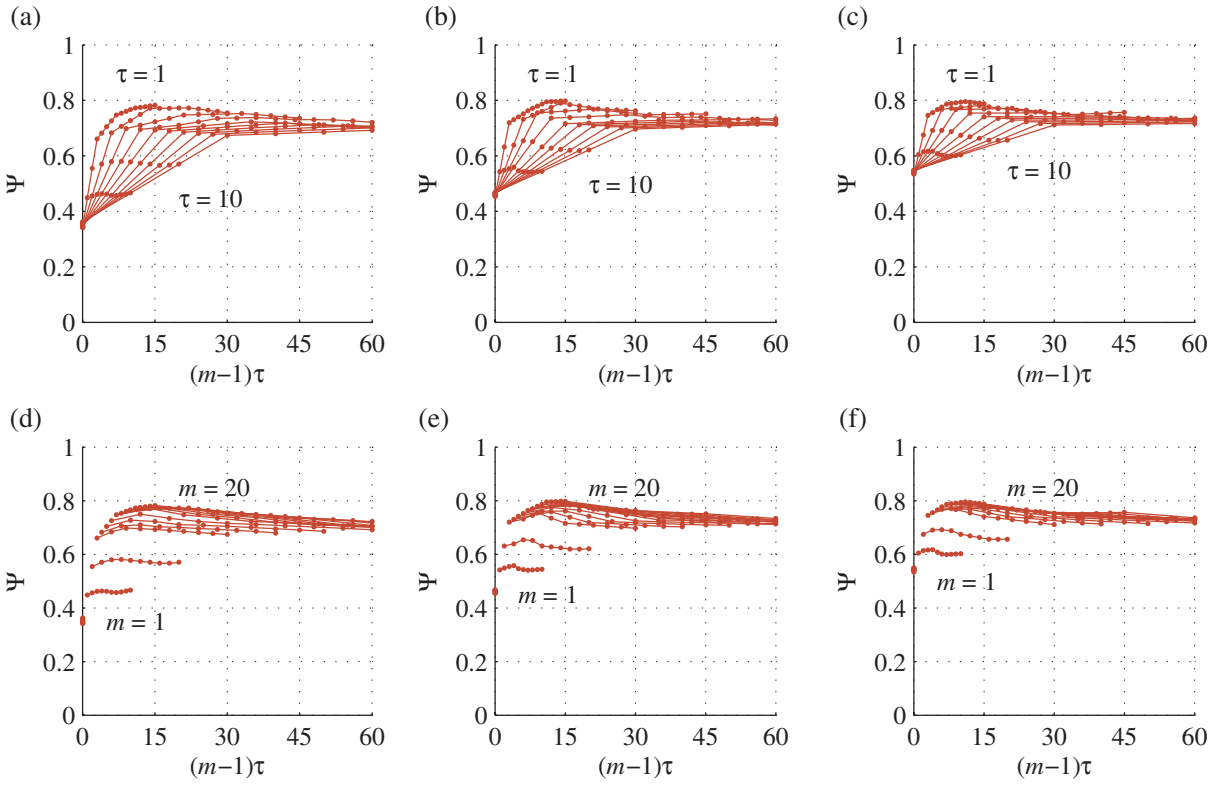

Figure 7. $Q_{\max}$ iso-τ (a)–(c) and iso-m (d)–(f) curves for $\kappa = 0.05$ (a,d), $\kappa = 0.1$ (b,e) and $\kappa = 0.15$ (c,f).

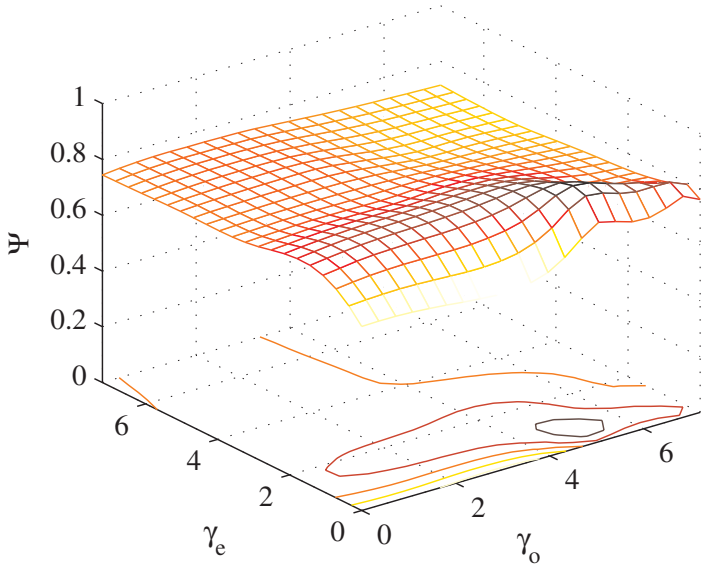

Figure 8. $\Psi_{Q_{\max}}$ in dependence on γ_o and γ_e values.

4.2. Out-of-sample accuracy

Accuracies for the testing collections using the parameters determined by the optimization on the training collection are shown in figures 9(b)–(d). Resulting average out-of-sample accuracies are $\Psi_{L_{\max}} = 0.426$, $\Psi_{S_{\max}} = 0.543$ and $\Psi_{Q_{\max}} = 0.667$. These good out-of-sample

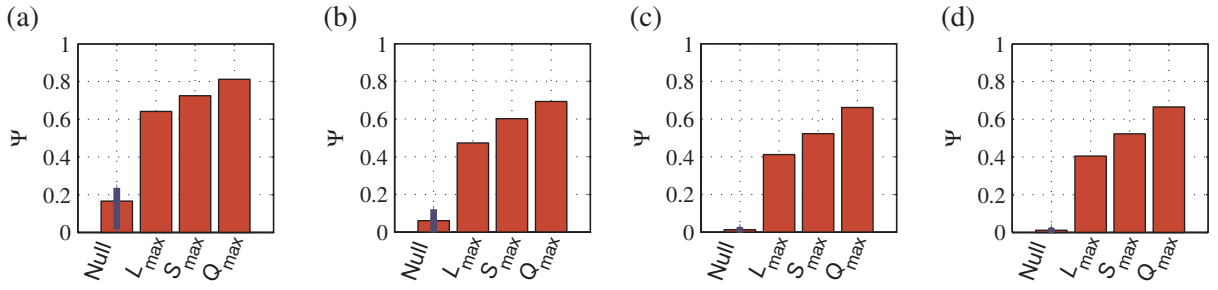

Figure 9. Mean average precision Ψ for the training (a) and the three testing collections (b)–(d). Error margins in the leftmost bars correspond to the range across to the 19 randomizations described in section 3.2.

accuracies indicate that our results cannot be explained by a parameter over-optimization. The accuracy increase gained through the derivation from $L_{\max}$ via $S_{\max}$ to $Q_{\max}$ is substantial. Most importantly, this increase in accuracy is reflected in the testing collections as well. Moreover, all values for $L_{\max}$, $S_{\max}$ and $Q_{\max}$ are significantly outside the range of Ψ_{null} across the 19 Monte Carlo randomizations. Therefore, our accuracy values are not consistent with the null hypothesis that the similarity matrices have no discriminative power.

4.3. Comparison with the state-of-the-art

As stated in the introduction, the algorithm proposed in [17] as well as two algorithms based on $Q_{\max}$ were submitted to the MIREX contest in 2007 and 2008, respectively. The MIREX test collection is composed of 30 cover sets of cardinality 11 each [20]. Accordingly, the total cover song collection contains 330 songs. Another 670 individual songs, i.e. cover sets of cardinality 1, are added to make the identification task more difficult. The entire music collection includes a wide diversity of genres (e.g. pop, rock, classical, baroque, folk, jazz, etc), and the variations span a variety of styles and orchestrations. Beyond this general description, no further information about the test collection is published or disclosed to the participants. In particular, only the MIREX organizers know what actual musical pieces are contained in the test collection. Each of the 330 cover songs were used as query and the submitted algorithms were required to return a 330 times 1000 distance matrix (one row for each query[5]). From this distance matrix, several evaluation measures were computed by the MIREX organizers. In 2007 and 2008 the same evaluation measures were applied, including Ψ as the main reference.

The algorithm in [17] was found to be the most accurate one in the 2007 edition[6] (figure 10(a), $\Psi_{[17]} = 0.521$). The two most accurate algorithms in 2008 were based on $Q_{\max}$. The raw $Q_{\max}$ algorithm as presented here reached an accuracy[7] of $\Psi_{Q_{\max}} = 0.661$ (figure 10(b)). It was only outperformed by an algorithm which included $Q_{\max}$ as described here, plus one additional simple post-processing step applied to the similarity matrix derived from $Q_{\max}$ ($\Psi_{Q^*_{\max}} = 0.750$). This post-processing step was proposed by our group and consists detecting cover song sets instead of isolated songs [55]. More concretely, it applies an unsupervised community detection algorithm operating to a complex network computed from

[5] http://www.music-ir.org/mirex/2008/index.php/Audio_Cover_Song_Identification
[6] http://www.music-ir.org/mirex/2007/index.php/Audio_Cover_Song_Identification_Results
[7] http://www.music-ir.org/mirex/2008/index.php/Audio_Cover_Song_Identification_Results

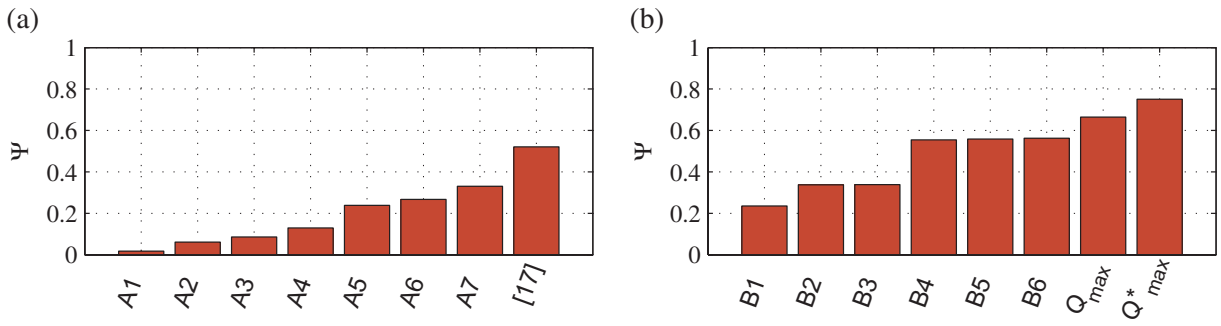

Figure 10. Mean average precision Ψ for algorithms submitted to the MIREX 2007 (a) and 2008 (b) contests. A1–A7 and B1–B6 refer to algorithms submitted by other participants and [17] refers to our previous work.

the pairwise $Q_{\max}$ matrix used in section 3.2 and normalize this matrix according to the detected communities.

Most importantly, the $\Psi_{Q_{\max}}$ value obtained for the MIREX music collection is very close to the $\Psi_{Q_{\max}}$ values reported for the testing collections used here (figures 9 and 10). This provides evidence that the out-of-sample accuracy values reported in section 4.2 are not related to any hidden in-sample optimization which could have been introduced involuntarily, for example, by a biased selection of songs for the testing collections.

5. Conclusion

In the present work, we combine concepts from music signal processing, nonlinear time series analysis, machine learning and IR to successfully identify covers of musical pieces. The composition of concepts from these different disciplines, naturally results in a modular organization of our method. Given two audio signals, we, at first, use techniques from music signal processing to extract descriptor time series representing their tonal progression. These time series are then used for multivariate embedding by means of delay coordinates. To assess equivalences of states between both systems attained at different times, we use CRPs and recurrence quantification measures derived from them. In pre-analysis, existing recurrence quantification measures were evaluated using machine learning techniques. The obtained result motivated us to introduce new cross recurrence quantification measures $S_{\max}$ and $Q_{\max}$. Using standard IR evaluation measures we quantify the accuracy for the task at hand.

We here show that our algorithm leads to high accuracy for the cover song identification task on a comprehensive music collection compiled prior to and independently from the present study. This music collection is divided into non-overlapping testing and training collections. We adjust the parameters on the training collection and then determine the accuracy out-of-sample using different testing collections. Nonetheless, in such a study design, one could still overestimate the true accuracy of the algorithm by involuntarily introducing biases in the compilation of the music collection. However, the close match of the accuracy reported here for our music collection and the one obtained for the MIREX contest supports the generality of the reported results (recall that the music collection used here was compiled prior to and independently from our participation to the MIREX contest). Furthermore, the proposed algorithm reached the highest accuracies in the MIREX cover song identification task ever. This

illustrates its superiority in respect to current state-of-the-art algorithms, including our previous approach [17].

One should note that the concept of delay coordinates has originally been developed for the reconstruction of stationary deterministic dynamical systems from single variables measured from them [22]. Also, the identification of coherent traces within the CRP is connected to the notion of deterministic dynamics (see [26] and references therein). Certainly, musical pieces do not represent the output of a stationary deterministic dynamical system, and therefore, one could argue that applying concepts developed for deterministic systems to such signals is inappropriate. However, if we consider a song as the output of some 'complicated system' evolving with time, and an HPCP as a multivariate time series measured from it, we can use the method of delay coordinates to facilitate the extraction of the information characterizing the underlying system. In fact, we find that the accuracy of our cover song identification system is significantly improved using an embedding, compared to not using it. In conclusion, our work provides a further example for an application of nonlinear time series analysis methods to experimental time series where the assumption of some underlying deterministic dynamics is not fulfilled in a strict sense, but which nonetheless allows one to successfully characterize the system underlying the time series.

6. Outlook

In closing, we provide evidence that the $Q_{\max}$ measure proposed here is not restricted to MIR nor to the particular application of cover song identification. Indeed, a quantitative assessment of curved and disrupted traces in RPs and CRPs can be useful for the characterization of a variety of experimental and artificial signals.

As a concrete example for a physical setting, we study two Rössler dynamics unidirectionally coupled by a diffusive term of strength ε:

$$\begin{aligned}
\dot{x}_1(t) &= -\omega_x(t)x_2(t) - x_3(t),\\
\dot{x}_2(t) &= \omega_x(t)x_1(t) + 0.15x_2(t),\\
\dot{x}_3(t) &= [x_1(t) - 10]\, x_3(t) + 0.2,\\
\dot{y}_1(t) &= -\omega_y y_2(t) - y_3(t) + \varepsilon\,[x_1(t) - y_1(t)]\,,\\
\dot{y}_2(t) &= \omega_y y_1(t) + 0.15y_2(t),\\
\dot{y}_3(t) &= [y_1(t) - 10]\, y_3(t) + 0.2.
\end{aligned} \tag{9}$$

For our context, the key feature of this example is that the mean frequency of the driving dynamics $\omega_x(t)$ is varied while $\omega_y = 1$ is time-independent. We integrate equation (9) using a fourth order Runge–Kutta algorithm with fixed step size of $\Delta t = 0.05$ time units and vary $\omega_x(t)$ according to

$$\omega_x(j\Delta t) = 1 + 0.02\xi_j, \tag{10}$$

where ξ_j is a strongly correlated first-order autoregressive process

$$\xi_j = 0.98\xi_{j-1} + \eta_j \tag{11}$$

with j being an integer and η_j corresponding to uncorrelated Gaussian noise with zero mean and unit variance. Note that ξ_j has zero mean and a variance of approximately 24. We start

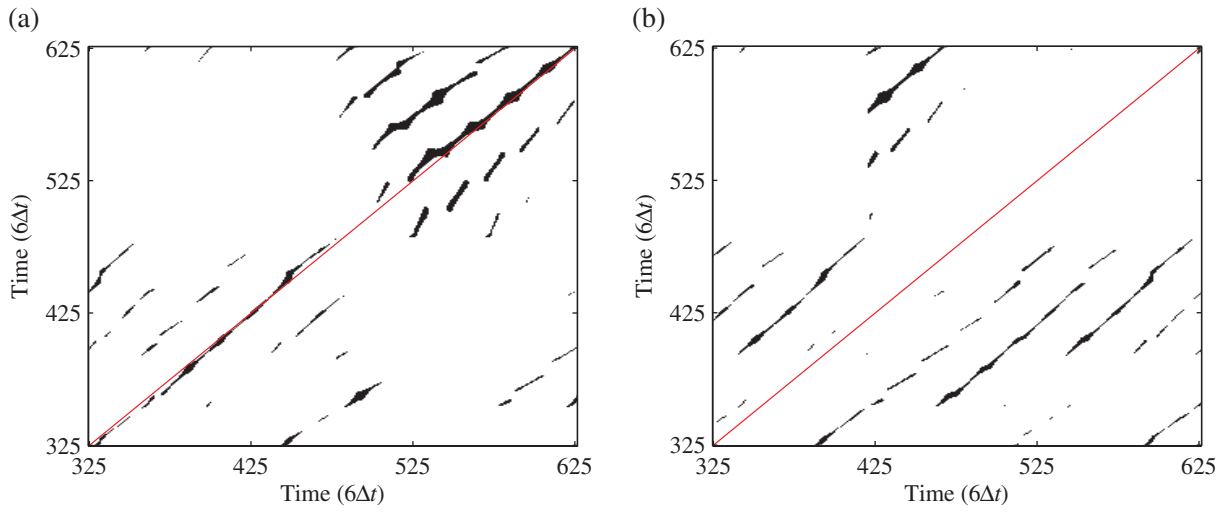

Figure 11. Exemplary CRP regions for $x_1(t_i)$ and $y_1(t_i)$ obtained from one realization of the coupled Rössler dynamics (a) and the uncoupled ones (b).

the integration of equation (9) at random initial conditions and use a sufficient number of pre-iterations to reduce transients. Time series pairs $x_1(t_i)$ and $y_1(t_i)$ are then sampled at $t_i - t_{i-1} = 6\Delta t$ for a time series length of $N_x^* = N_y^* = 2048$ ($i = 1, \ldots, 2048$).

We compare results for coupled dynamics (equation (9) with $\varepsilon = 0.4$) versus uncoupled dynamics (equation (9) with $\varepsilon = 0$). For both conditions, we generate a set of 2000 independent realizations for each time series $x_1(t_i)$ and $y_1(t_i)$. We construct the CRP and extract $L_{\max}$, $S_{\max}$ and $Q_{\max}$ from all time series pairs. We here use as parameters: $m = 8$, $\tau = 1$, $\kappa = 0.0125$ and $\gamma_o = \gamma_e = 1$. None of these parameters, nor the parameters of equations (9)–(11), are optimized in any way for the example presented here.

Regarding the CRP constructed from realizations of $x_1(t)$ and $y_1(t)$ for the coupled dynamics, we find curved and briefly disrupted traces along the main diagonal (figure 11(a)). These reflect the strong coupling and their interruptions and curvatures are caused by the stochastically varying mean frequency of the driving Rössler oscillator. In contrast, only dispersed patterns are observed for the uncoupled dynamics (figure 11(b)). In consequence, across all realizations, the distributions of $Q_{\max}$ values obtained for the coupled versus uncoupled condition are almost non-overlapping (figure 12(c)). Distributions of $L_{\max}$ and $S_{\max}$ in contrast overlap substantially (figures 12(a) and (b), respectively). Hence, only $Q_{\max}$ allows one to distinguish between these two conditions.

This example of coupled Rössler dynamics with stochastically varying mean frequencies is meant to sketch only one potential application of $Q_{\max}$. A systematic study of this setting and the influence of the various parameters is left for future work. Results of such a study can have important implications for the analysis of interactions between brain oscillations and tremors in Parkinson patients or between cardiac and respiratory dynamics. This holds since these pathological and physiological processes are known to be characterized by mean frequencies with irregular time-dependencies.

Furthermore, one should note that curved structures have been reported in RPs and CRPs of artificial and experimental signals. Artificial signals include frequency modulated periodic signals [29, 30, 56] or time series derived from Rössler dynamics with bidirectional couplings close to the onset of phase synchronization [56]. Experimental data include signals with

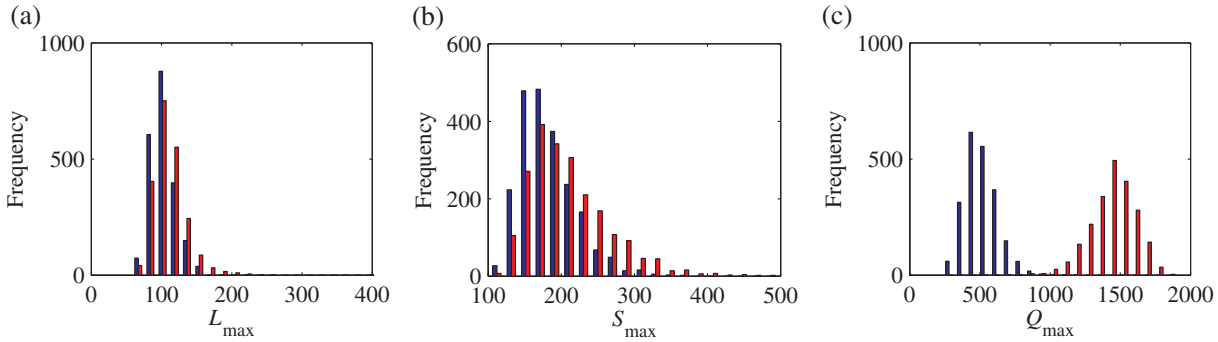

Figure 12. Histograms for L_{max} (a), S_{max} (b) and Q_{max} (c) obtained for the 2000 independent realizations of the coupled (red) and uncoupled (blue) dynamics.

nonlinearly re-scaled or distorted time axes such as geophysical data of sediment cores subjected to different compressions [29], symbolic dynamic representations of EEG recordings from the onsets of epileptic seizures [56], or acoustic signals from calls of primates [30]. Far beyond these particular examples, it can be conjectured that important features of further experimental signals, e.g. from bioinformatics [57], physiology [24], human speech processing [51], or climatology [58], are reflected in curved and disrupted traces in RPs and CRPs. A quantitative assessment of these traces by means of the proposed measures can thus help to characterize a multitude of systems from different scientific disciplines.

Acknowledgments

We thank D Chicharro, E Gómez and P Herrera for useful discussions and M Koppenberger for technical support. This research has been partially funded by the EU-IP project PHAROS IST-2006-045035 and by the BFU2007-61710 Spanish Ministry of Education and Science grant.

References

[1] Casey M, Veltkamp R C, Goto M, Leman M, Rhodes C and Slaney M 2008 Content-based music information retrieval: current directions and future challenges *Proc. IEEE* **96** 668–96
[2] Orio N 2006 Music retrieval: a tutorial and review *Found. Trends Inf. Retrieval* **1** 1–90
[3] Tzanetakis G and Cook P 2002 Musical genre classification of audio signals *IEEE Trans. Speech Audio Process.* **5** 293–302
[4] Aucouturier J J and Pachet F 2004 Improving timbre similarity: how high is the sky? *J. Negative Results Speech Audio Sci.* **1** 1
[5] Bergstra J, Casagrande N, Erhan D, Eck D and Kégl B 2006 Aggregate features and adaboost for music classification *Mach. Learn. J.* **65** 473–84
[6] Müller M 2007 *Information Retrieval for Music and Motion* (Berlin: Springer)
[7] Ong B S 2007 Structural analysis and segmentation of music signals *PhD Thesis* Universitat Pompeu Fabra, Barcelona, Spain Available online: http://mtg.upf.edu/node/508
[8] Casey M, Rhodes C and Slaney M 2008 Analysis of minimum distances in high-dimensional musical spaces *IEEE Trans. Audio Speech Lang. Process.* **16** 1015–28
[9] Cano P, Batlle E, Kalker T and Haitsma J 2005 A review of audio fingerprinting *J. VLSI Signal Process.* **41** 271–84

[10] Nagano H, Kashino K and Murase H 2002 Fast music retrieval using polyphonic binary feature vectors *IEEE Int. Conf. on Multimedia and Expo (ICME)* vol 1, pp 101–4
[11] Tsai W H, Yu H M and Wang H M 2008 Using the similarity of main melodies to identify cover versions of popular songs for music document retrieval *J. Inf. Sci. Eng.* **24** 1669–87
[12] Izmirli Ö 2005 Tonal similarity from audio using a template based attractor model *Int. Symp. on Music Information Retrieval (ISMIR)* pp 540–5
[13] Gómez E, Ong B S and Herrera P 2006 Automatic tonal analysis from music summaries for version identification *Conv. of the Audio Engineering Society (AES)* CD-ROM, paper no. 6902
[14] Marolt M 2008 A mid-level representation for melody-based retrieval in audio collections *IEEE Trans. Multimedia* **10** 1617–25
[15] Ellis D P W and Poliner G E 2007 Identifying cover songs with chroma features and dynamic programming beat tracking *IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)* vol 4, pp 1429–32
[16] Bello J P 2007 Audio-based cover song retrieval using approximate chord sequences: testing shifts, gaps, swaps and beats *Int. Symp. on Music Information Retrieval (ISMIR)* pp 239–44
[17] Serrà J, Gómez E, Herrera P and Serra X 2008 Chroma binary similarity and local alignment applied to cover song identification *IEEE Trans. Audio Speech Lang. Process.* **16** 1138–52
[18] Serrà J, Gómez E and Herrera P 2009 *Audio Cover Song Identification and Similarity: Background, Approaches, Evaluation, and Beyond* (Berlin: Springer) at press
[19] Jensen J H, Christensen M G, Ellis D P W and Jensen S H 2008 A tempo-insensitive distance measure for cover song identification based on chroma features *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* pp 2209–12
[20] Downie J S, Bay M, Ehmann A F and Jones M C 2008 Audio cover song identification: MIREX 2006–2007 results and analyses *Int. Symp. on Music Information Retrieval (ISMIR)* pp 468–73
[21] Downie J S 2008 The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research *Acoust. Sci. Technol.* **29** 247–55
[22] Kantz H and Schreiber T 2004 *Nonlinear Time Series Analysis* 2nd edn (Cambridge: Cambridge University Press)
[23] Zbilut J P and Webber C L Jr 1992 Embeddings and delays as derived from quantification of recurrence plots *Phys. Lett.* A **171** 199–203
[24] Webber C L Jr and Zbilut J P 1994 Dynamical assessment of physiological systems and states using recurrence plot strategies *J. Appl. Physiol.* **76** 965–73
[25] Marwan N, Wessel N, Meyerfeldt U, Schirdewan A and Kurths J 2002 Recurrence-plot-based measures of complexity and its application to heart rate variability data *Phys. Rev.* E **66** 026702
[26] Marwan N, Romano M C, Thiel M and Kurths J 2007 Recurrence plots for the analysis of complex systems *Phys. Rep.* **438** 237–329
[27] Zbilut J P, Giuliani A and Webber C L Jr 1998 Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification *Phys. Lett.* A **246** 122–8
[28] Eckmann J P, Kamphorst S O and Ruelle D 1987 Recurrence plots of dynamical systems *Europhys. Lett.* **5** 973–7
[29] Marwan N, Thiel M and Nowaczyk N R 2002 Cross recurrence plot based synchronization of time series *Nonlinear Process. Geophys.* **9** 325–31
[30] Facchini A, Kantz H and Tiezzi E 2005 Recurrence plot analysis of nonstationary data: the understanding of curved patterns *Phys. Rev.* E **72** 021915
[31] Gerhard D 1999 Audio visualization in phase space *Bridges: Mathematical Connections in Art, Music, and Science* pp 137–44
[32] Reiss J D and Sandler M B 2003 Nonlinear time series analysis of musical signals *Int. Conf. on Digital Audio Effects (DAFx)* pp 1–5
[33] Mierswa I and Morik K 2005 Automatic feature extraction for classifying audio data *Mach. Learn. J.* **58** 127–49

[34] Mörchen F, Ultsch A, Thies M and Löhken I 2006 Modelling timbre distance with temporal statistics from polyphonic music *IEEE Trans. Speech Audio Process.* **14** 81–90
[35] Mörchen F, Mierswa I and Ultsch A 2006 Understandable models of music collections based on exhaustive feature generation with temporal statistics *ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 882–91
[36] Hegger R, Kantz H and Matassini L 2000 Denoising human speech signals using chaoslike features *Phys. Rev. Lett.* **84** 3197–200
[37] Matassini L, Kantz H, Holyst J and Hegger R 2002 Optimizing of recurrence plots for noise reduction *Phys. Rev.* E **65** 021102
[38] Foote J 1999 Visualizing music and audio using self-similarity *ACM Int. Conf. on Multimedia* pp 77–80
[39] Foote J 2000 Automatic audio segmentation using a measure of audio novelty *IEEE Int. Conf. on Multimedia and Expo (ICME)* vol 1452–55
[40] Casey M and Westner W 2000 Separation of mixed audio sources by independent subspace analysis *Int. Computer Music Conf. (ICMC)* pp 154–61
[41] Gainza M 2009 Automatic musical meter detection *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* pp 329–32
[42] Fujishima T 1999 Realtime chord recognition of musical sound: a system using common lisp music *Int. Computer Music Conference (ICMC)* pp 464–7
[43] Sheh A and Ellis D P W 2003 Chord segmentation and recognition using EM-trained hidden Markov models *Int. Symp. on Music Information Retrieval (ISMIR)* pp 183–9
[44] Paws S 2004 Musical key extraction from audio *Int. Symp. on Music Information Retrieval (ISMIR)* pp 96–9
[45] Gómez E 2006 Tonal description of music audio signals *PhD Thesis*, Universitat Pompeu Fabra, Barcelona, Spain. Available online: http://mtg.upf.edu/node/472.
[46] Serrà J, Gómez E and Herrera P 2008 Transposing chroma representations to a common key *IEEE CS Conf. on The Use of Symbols to Represent Music and Multimedia Objects* pp 45–8
[47] Takens F 1981 Detecting strange attractors in turbulence *Lect. Notes Math.* **898** 366–81
[48] Huron D 2006 *Sweet Anticipation: Music and the Psychology of Expectation* (Cambridge: MIT Press)
[49] Schulkind M D, Posner R J and Rubin D C 2003 Musical features that facilitate melody identification: how do you know it's your song when they finally play it? *Music Percep.* **21** 217–49
[50] Witten I H and Frank E 2005 *Data Mining: Practical Machine Learning Tools and Techniques* 2nd edn (Amsterdam: Elsevier)
[51] Rabiner L R and Juang B H 1993 *Fundamentals of Speech Recognition* (New York: Prentice-Hall)
[52] Gusfield D 1997 *Algorithms on Strings, Trees and Sequences: Computer Sciences and Computational Biology* (Cambridge: Cambridge University Press)
[53] Myers C, Rabiner L R and Rosenberg A E 1980 Performance tradeoffs in dynamic time warping algorithms for isolated word recognition *IEEE Trans. Audio Speech Lang. Process.* **28** 623–35
[54] Manning C D, Prabhakar R and Schutze H 2008 *An Introduction to Information Retrieval* (Cambridge: Cambridge University Press) Available online: http://www.informationretrieval.org.
[55] Serrà J, Zanin M, Laurier C and Sordo M 2009 Unsupervised detection of cover song sets: accuracy increase and original detection *Conf. of the Int. Society for Music Information Research (ISMIR)* at press
[56] Groth A 2005 Visualization of coupling in time series by order recurrence plots *Phys. Rev.* E **72** 046220
[57] Aach J and Church G 2001 Aligning gene expression time series with time warping algorithms *Bioinformatics* **17** 495–508
[58] Marwan N and Kurths J 2002 Nonlinear analysis of bivariate data with cross recurrence plots *Phys. Lett.* A **302** 299–307