Predictability of Music Descriptor Time Series and its Application to Cover Song Detection

Joan Serrà, Holger Kantz, Xavier Serra and Ralph G. Andrzejak

Abstract

Intuitively, music has both predictable and unpredictable components. In this work we assess this qualitative statement in a quantitative way using common time series models fitted to state-of-the-art music descriptors. These descriptors cover different musical facets and are extracted from a large collection of real audio recordings comprising a variety of musical genres. Our findings show that music descriptor time series exhibit a certain predictability not only for short time intervals, but also for mid-term and relatively long intervals. This fact is observed independently of the descriptor, musical facet and time series model we consider. Moreover, we show that our findings are not only of theoretical relevance but can also have practical impact. To this end we demonstrate that music predictability at relatively long time intervals can be exploited in a real-world application, namely the automatic identification of cover songs (i.e. different renditions or versions of the same musical piece). Importantly, this prediction strategy yields a parameter-free approach for cover song identification that is substantially faster, allows for reduced computational storage and still maintains highly competitive accuracies when compared to state-of-the-art systems.

EDICS Category: AUD-CONT

This work has been partially funded by the *Deutscher Akademischer Austausch Dienst* (A/09/96235), by the projects Classical Planet (MITYC: TSI-070100- 2009-407) and DRIMS (MICINN: TIN2009-14247-C02-01), by the Spanish Ministry of Education and Science (BFU2007-61710) and by the Max Planck Institute for the Physics of Complex Systems.

J. Serrà, X. Serra and R. G. Andrzejak are with Universitat Pompeu Fabra, Roc Boronat 138, 08018 Barcelona, Spain, phone +34 93 542 2864, fax +34 93 542 2517 (e-mail: joan.serraj@upf.edu, xavier.serra@upf.edu, ralph.andrzejak@upf.edu). H. Kantz is with the Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Strasse 38, 01187 Dresden, Germany, phone +49 351 871 2216, fax +49 351 871 1999 (e-mail: kantz@pks.mpg.de).

Predictability of Music Descriptor Time Series and its Application to Cover Song Detection

I. Introduction

Music is ubiquitous in our lives and we have been enjoying it since the beginning of human history [1]. This enjoyment of music is intrinsically related to the ability to anticipate forthcoming events [2]–[4]. Indeed, the meeting of our musical expectations plays a fundamental role in the aesthetics and emotions of music. Yet, the secret of a good song remains in the right balance between predictability and surprise. It is this balance that makes music interesting for us. Accordingly, we tend to dislike music that is either extremely simple or extremely complex [2]–[4]. Paraphrasing Levitin [2], we could say that the act of listening to music "rewards us for correct predictions" but, at the same time, "challenges us with new organizational principles". Thus music seems to be intrinsically predictable and unpredictable at the same time: we know it has some (predictable) structures and repetitions, but we can also state that there is a strong random (unpredictable) component.

Although very intuitive, the above dichotomy has scarce quantitative evidence. The majority of quantitative studies have been conducted with music scores or symbolic note transcriptions of selected, usually Western classical compositions [3]–[8]. Some of them just consider melodic, simple, synthetic and/or few musical examples [6]–[8]. Thus the question arises if quantitative evidence is found in large-scale corpora of music including different geographical locations and genres apart from Western classical music. Furthermore, there is a lack of knowledge with regard to the predictability of musical facets other than melodic or harmonic ones, e.g. timbre, rhythm or loudness. And, what is even more surprising, there are only few experiments with real recordings [9]–[11]. Such consideration is important, since scores or symbolic representations do not faithfully reflect primary perceptual elements of musical performance and expressiveness, which are in turn related to our sensation of surprise and to the predictability of the piece. Finally, existing studies usually restrict the analyzes to the transitions between consecutive elements. Hence they do not consider different prediction intervals or *horizons* (i.e. how far in the future we perform predictions). This fact raises further questions: How does such horizon affect the predictability of music? Do different musical facets exhibit a similar predictability at short as well as long time intervals? How do these predictabilities behave in dependence of the prediction horizon? All these questions are important to advance towards a better scientific understanding of music (c.f. [1]–[11]).

The questions above can be addressed by using tools from music information retrieval (MIR) [12]–[15]. MIR is an interdisciplinary research field that aims at automatically understanding, describing, retrieving and organizing musical content. In particular, much effort is focused on extracting qualitative and quantitative information from the audio signal of real recordings in order to represent certain musical aspects such as timbre, rhythm, melody, main tonality, chords or tempo [12], [14]. Quantitative descriptions of such aspects are computed in a short-time moving window either from a temporal, spectral or cepstral representation of the audio signal [15]. This computation leads to a time series reflecting the temporal evolution of a given musical facet: a *music descriptor time series*.

Music descriptor time series are essential for quantitative large-scale studies on the predictability of real recordings. Indeed, when assessing the predictability of such time series, we are assessing the predictability of the musical facet they represent. A straightforward way to perform this assessment is to fit a model to the time series and to evaluate the in-sample *self-prediction* error of the model forecasts. If similar results are observed for a variety of model classes, we can then conclude that what we observe is, in fact, a direct product of the information conveyed by the time series and not an artifact of the particular model being employed. In a similar manner, if we work with different descriptor time series representing the same musical facet, we can be more confident that what we see is due to the musical facet and not to the particular descriptor used.

In the present work we therefore study a variety of different descriptor time series reflecting complementary musical facets. These descriptors are extracted from a large collection of real recordings covering multiple genres. Furthermore, we consider a number of simple time series models, the predictions of which are studied for a range of horizons. Our analysis unveils that a certain predictability is observed for a broad range of prediction horizons. While absolute values of the prediction errors vary, we find a number of general features across models, descriptor time series and musical facets. In particular, we show that the error in the predictions, in spite of being high and rapidly increasing at short horizons, saturates at values lower than expected for random data at medium and relatively long time intervals. Furthermore, we note that this error grows sub-linearly, following a square root or logarithmic curve.

Together with these advances in scientific knowledge, we provide a direct real-world application of music prediction concepts, namely the automatic detection of *cover songs* (i.e. different renditions of the same musical composition). To this end, we use out-of-sample *cross-predictions*. That is, we train a model with a time series of one recording and then use this model to predict the time series of a different recording. Intuitively, once a model has learned the patterns found in the time series of a given query song, one should expect the average prediction error to be relatively small when the time series of

a candidate cover song is used as input. Otherwise, i.e. when an unrelated (non-cover) candidate song is considered, the prediction error should be higher. Indeed, we demonstrate that such a model-based cross-prediction strategy can be effectively employed to automatically detect cover songs.

Cover song detection has been a very active area of study within the MIR community over the last years [16]. This is due to the introduction of digital ways to share and distribute information, which represent a challenge for the search and organization of musical contents [13]–[15]. Cover song detection (or identification) is a very simple task from the user's perspective: a query song is provided, and the system is asked to retrieve all versions of it in a given music collection. However, from an MIR perspective it becomes a very challenging task, since cover songs can differ from their originals in several musical aspects such as timbre, tempo, song structure, main tonality, arrangement, lyrics or language of the vocals [16]. In spite of these differences, cover songs might retain a considerable part of their tonal progression (e.g. melody and/or chords). Hence the large majority of state-of-the-art approaches are based on the detection of common patterns in the dynamics of tonal descriptors (e.g. [17]–[20]). Another major characteristic that is shared among all state-of-the-art approaches for cover song detection is the lack of specific modeling strategies for descriptor time series [16]. This is somehow surprising regarding the benefits arising from the generality and the compactness of the description. Indeed, modeling strategies have been successfully employed in a number of similar tasks such as exact audio matching [15], in other MIR problems [12] or in related areas such as speech processing [21].

In the present work we show that a model-based forecasting strategy for cover song detection is very promising in the sense that it achieves competitive accuracies and it provides advantages when compared to state-of-the-art approaches, such as lower computational complexities and potentially fewer storage requirements. But perhaps the most interesting aspect of such strategy is that no parameters need to be adjusted. More specifically, model parameters and coefficients are automatically learned for each song and descriptor time series individually. No intervention of the user is needed. Accordingly, the system can be readily applied to different music collections or descriptor time series.

The rest of the paper is organized as follows. We first present an overview of our methodology, including specific details of the employed music descriptor time series and the considered models (Sec. II). We then explain our evaluation measures and data (Sec. III). The results section follows, both on the self-prediction (i.e. predictability assessment) and on cross-prediction (i.e. cover song detection) experiments (Sec. IV). A discussion of our model-based strategy for cover song detection is provided (Sec. V) before we summarize our main findings and outline some future perspectives (Sec. VI).

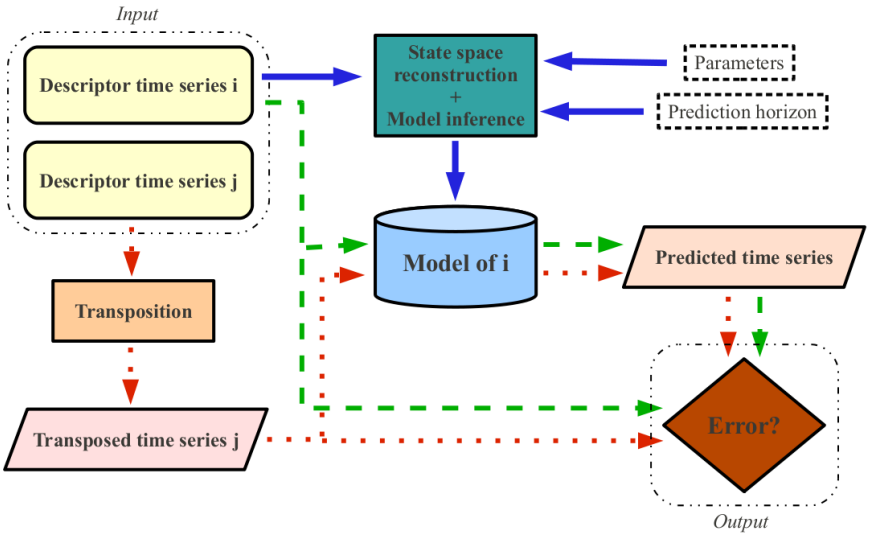

Fig. 1. Schematic diagram for self- and cross-prediction experiments. Broken green arrows correspond to the self-prediction experiments and dotted red arrows correspond to the cross-prediction ones. Solid blue arrows correspond to the training phase for both self- and cross-prediction experiments.

II. Methodology

A. Overview

Our analysis follows a typical modeling and forecasting architecture (Fig. 1). In the case of self-prediction, we assess the error produced by a model trained on descriptor time series i when the same time series i is used as input (Fig. 1, broken green arrows). In particular, this training consists of the determination of optimal model parameters and coefficients. In the case of cross-prediction, we are interested in the error produced by a model trained on descriptor time series i when a different time series j is used as input (Fig. 1, dotted red arrows). This cross-prediction error is then taken as a pairwise dissimilarity measure between cover songs (see below). The training of the models is done in the same way for self- and cross-prediction experiments (Fig. 1, solid blue arrows). All the aforementioned processes are carried out for a range of prediction horizons h.

Before we can exploit the concept of cross-prediction errors as a dissimilarity measure, we have to investigate predictability itself and establish a set of methods (or models) for descriptor prediction. As input for the models we consider twelve standard descriptor time series reflecting complementary musical facets related to tonality, timbre, rhythm and loudness. The specific descriptors considered are: pitch class profiles, tonal centroid, harmonic change, Mel frequency cepstral coefficients, spectral contrast, spectral

peaks, spectral harmonicity, onset probability, rhythm transform, onset coefficients, energy and loudness (Sec. II-B). As time series models we employ five common, simple time series models, both linear and nonlinear: autoregressive, threshold autoregressive, radial basis functions, locally constant and naïve Markov (Sec. II-C).

The cross-prediction error provides an estimation of the dissimilarity between two time series. This fact is exploited to detect cover songs. As mentioned, the only musical characteristic which is largely preserved across different cover versions is the tonal sequence. Therefore, for cross-prediction, we only consider tonal descriptor time series (pitch class profiles, tonal centroid and harmonic change). Importantly, we use transposed tonal descriptor time series. Cover versions may be played in different tonalities (e.g. to be adapted to the characteristics of a particular singer or instrument), and this changes might be reflected in the tonal descriptor time series j. To counteract this effect, various strategies have been devised in the literature [16]. In particular, we here use the so-called optimal transposition index method [22], which is applied to time series j before the forecasting process (Fig. 1, bottom left). This method is commonly used in a variety of cover song identification systems [16], including our previous work [19], [20]. For a detailed explanation and evaluation of this transposition method we refer to [22].

B. Music descriptor time series

We use a set of twelve state-of-the-art descriptors reflecting the dynamics of complementary musical aspects related to tonality, timbre, rhythm and loudness. Some of these descriptors are computed with an in-house tool specifically designed for that purpose [23], while for others we implement the algorithms from the cited references. Our focus is not in the descriptor extraction process itself, but in studying the predictability of different musical facets represented by common descriptor time series. Therefore, for the sake of brevity, we skip the underlying mathematical formulae and details of numerical algorithms and refer to the citations below.

- Tonality: Within this musical facet we consider pitch class profiles (PCP) [24], tonal centroid (TC), and harmonic change (HC) descriptors [25]. The first two represent the tonal content of the song, i.e. the content related to the relative energy of the different notes of the Western chromatic scale. The third one represents the degree of change of tonal content between successive windows. Except for the fact that we employ 12-dimensional vectors, the extraction process for PCPs is the same as in [19]. Once PCPs are obtained, deriving TC and HC is straightforward [25].
- Timbre: Mel frequency cepstral coefficients (MFCC) are routinely employed in MIR. We use the Auditory toolbox implementation [26] with 12 coefficients (skipping the DC coefficient). Apart from

MFCCs we also consider the spectral contrast (SC) [27], the spectral peaks (SP) and the spectral harmonicity (SH) descriptors [23]. SC is 12-dimensional, SP and SH are unidimensional. MFCCs and SCs collect general (or global) timbre characteristics. SP quantifies the number of peaks in a window's spectrum and SH quantifies how these peaks are distributed.

- Rhythm: We use an onset probability (OP) curve [28] denoting, for each analysis window, the likelihood of the beginning of a musical note. We also employ more broad representations of rhythm content like the rhythm transform (RT) [29] or the onset coefficients (OC) [30]. These two descriptors are computed using a total window length of 6 s, a much longer window than the one used for other descriptors (see below). This is because in order to obtain a general representation of rhythm, a longer time span must be considered (e.g. no proper rhythm can be subjectively identified by a listener in 100 or even 500 ms) [31]. For RT and OC, a final discrete cosine transform is used to compress the information. We use 20 coefficients for the former and 30 for the latter (DC coefficients are discarded).
- Loudness: Two unidimensional descriptors related to the overall loudness are used. One is simply the total energy (E) of the power spectrum of the audio window [32]. The other is a common loudness (L) descriptor [33], i.e. a psychological correlate of the auditory sensation of acoustic strength.

All considered descriptors are extracted from spectral representations of the raw audio signal in a moving window (frame by frame analysis). We use a step size of 116 ms and, unless stated otherwise, a window length of 186 ms[1]. Therefore, our resulting descriptor time series have a sampling rate of approximately 9.6 Hz (e.g. a song of 4 minutes yields a music descriptor time series of 2304 samples). These samples can be unidimensional or multidimensional, depending on the considered descriptor. We denote multidimensional descriptor time series as a matrix $\mathcal{S} = [\mathbf{s}_1 \dots \mathbf{s}_N]$, where N is the total number of samples and $\mathbf{s}_n$ is a column vector with D components representing a D-dimensional descriptor at sample window n. Therefore, element $s_{d,n}$ of $\mathcal{S}$ represents the magnitude of the d-th descriptor component of the n-th window.

C. Time series models

All the models described hereafter aim at the prediction of future states of dynamical systems based on their present states [34]. However, the information about the present state is, in general, not fully

[1]We initially extract descriptors for 93 ms frames (4096 samples at 44100 Hz) with 75% overlap and then average in blocks of 5 consecutive frames.

contained in a single sample from a time series measured from the dynamical system. To achieve a more comprehensive characterization of the present state one can take into account samples from the recent past. This is formalized by the concept of time delay embedding [35], also termed delay coordinate state space reconstruction. In our case, for multidimensional samples $\mathbf{s}_n$, we construct delay coordinate state space vectors $\mathbf{s}_n^*$ through vector concatenation, i.e.

$$\mathbf{s}_n^* = \begin{pmatrix} \mathbf{s}_n^{\mathrm{T}} & \mathbf{s}_{n-\tau}^{\mathrm{T}} & \cdots & \mathbf{s}_{n-(m-1)\tau}^{\mathrm{T}} \end{pmatrix}^{\mathrm{T}}, \tag{1}$$

where superscript $^{\mathrm{T}}$ denotes vector transposition, m is the embedding dimension and τ is the time delay. The sequence of these reconstructed samples yields again a multidimensional time series $\mathcal{S}^* = \left[\mathbf{s}_{w+1}^* \ldots \mathbf{s}_N^*\right]$, where $w = (m-1)\tau$ corresponds to the so-called embedding window. Notice that Eq. (1) still allows for the use of the raw time series samples (i.e. if $m = 1$ then $\mathcal{S}^* = \mathcal{S}$).

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 [35]. Certainly, a music descriptor time series does not represent a signal measured from a stationary dynamical system which could be described by some equation of motion. Nonetheless, delay coordinates, a tool that is routinely used in nonlinear time series analysis [34], can be pragmatically employed to facilitate the extraction of information contained in $\mathcal{S}$. In particular, the reconstruction of a state space by means of delay coordinates allows us to join the information about current and previous samples. Noticeably, there is evidence that such reconstruction can be beneficial for music retrieval [20], [36], [37].

To model and forecast music descriptor time series we employ popular, simple, yet flexible time series models; both linear and nonlinear [34], [38]–[41]. Since we do not have a good and well-established model for music descriptor prediction, we try a number of standard tools in order to identify the most suitable one. All modeling approaches we employ have clearly different features. Therefore they are able to exploit, in a forecasting scenario, different structures that might be found in the data. As linear approach we consider autoregressive models. Nonlinear approaches include locally constant, locally linear, globally nonlinear and probabilistic predictors.

For the majority of the following approaches we use a partitioning algorithm to divide a space into representative clusters. For this purpose we use a reimplementation of the K-medoids algorithm from [42]. The K-medoids algorithm [43] is a partitional clustering algorithm that attempts to minimize the distance between the points belonging to a cluster and the center of this cluster. The procedure to obtain the clusters is the same as with the well-known K-means algorithm [43] but, instead of using the mean

of the elements in a cluster, the medoid[2] is employed. The K-medoids algorithm is more robust to noise and outliers than the K-means algorithm [42], [43]. Usually, these two algorithms need to be run several times in order to achieve a reliable partition. However, the algorithm we use [42] incorporates a novel method for assigning the initial medoid seeds, which results in a deterministic and (most of the times) optimal cluster assignation.

1) Autoregressive (AR): A widespread way to model linear time series data is through an AR process, where predictions are based on a linear combination of m previous measurements [38]. We here employ a multivariate AR model [39]. In particular, we first construct delay coordinate state space vectors $\mathbf{s}_n^*$ and then express the forecast $\hat{\mathbf{s}}_{n+h}$ at h steps ahead from the n-th sample $\mathbf{s}_n$ as

$$\hat{\mathbf{s}}_{n+h} = \mathcal{A}\ \mathbf{s}_n^*, \tag{2}$$

where $\mathcal{A}$ is the $D \times mD$ coefficient matrix of the multivariate AR model. By considering samples $n = w+1, \ldots N-h$, one obtains an overdetermined system

$$\hat{\mathcal{S}} = \mathcal{A}\ \mathcal{S}^* \tag{3}$$

which, by ordinary least squares fitting [44], allows to estimate the matrix $\mathcal{A}$. It should be noticed that AR models have been previously used to characterize music descriptor time series in genre and instrument classification tasks [45], [46].

2) Threshold autoregressive (TAR): TAR models generalize AR models by introducing nonlinearity [47]. A single TAR model consists of a collection of AR models where each single one is valid only in a certain domain of the reconstructed state space (separated by the "thresholds"). This way, points in state space are grouped into patches and each of these patches is used to determine the coefficients of a single AR model (piecewise linearization).

For determining all TAR coefficients we partition the reconstructed space formed by $\mathcal{S}^*$ into K non-overlapping clusters with a K-medoids algorithm [42] and determine, independently for each partition, AR coefficients as above [Eqs. (2,3)]. Importantly, each of the K AR models is associated to the corresponding cluster. When forecasting, we again construct delay coordinate state space vectors $\mathbf{s}_n^*$ from each input sample $\mathbf{s}_n$, calculate their squared Euclidean distance to all $k = 1, \ldots K$ cluster medoids, and forecast

$$\hat{\mathbf{s}}_{n+h} = \mathcal{A}^{(k')}\ \mathbf{s}_n^*, \tag{4}$$

[2]A medoid is the representative item of a cluster whose average dissimilarity to all cluster items is minimal. In analogy to the median, the medoid has to be an existing element inside the cluster.

where $\mathcal{A}^{(k')}$ is the $D \times mD$ coefficient matrix of the multivariate AR model associated to the cluster whose medoid is closest to $\mathbf{s}_n^*$.

3) Radial basis functions (RBF): A very flexible class of global nonlinear models are RBF [48]. As with TAR, one partitions the reconstructed state space into K clusters but, in contrast, a scalar RBF function $\phi(x)$ is used for forecasting such that

$$\hat{\mathbf{s}}_{n+h} = \mathbf{b}_0 + \sum_{k=1}^{K} \mathbf{b}_k \; \phi\left(\|\mathbf{s}_n^* - \mathbf{c}_k\|\right), \tag{5}$$

where $\mathbf{b}_k$ are coefficient vectors, $\mathbf{c}_k$ are the cluster centers and $\| \; \|$ is some norm. In our case we use the cluster medoids for $\mathbf{c}_k$, the Euclidean norm for $\| \; \|$ and a Gaussian RBF function

$$\phi(x) = e^{-\frac{x^2}{2\alpha\rho_k}}. \tag{6}$$

We partition the space formed by $\mathcal{S}^*$ with the K-medoids algorithm, set ρ_k to the mean distance found between the elements inside the k-th cluster and leave α as a parameter. Notice that for fixed centers $\mathbf{c}_k$ and parameters ρ_k and α, determining the model coefficients becomes a linear problem that can be resolved again by ordinary least squares minimization. Indeed, a particularly interesting remark about RBF models is that they can be viewed as a (nonlinear, layered, feed-forward) neural network where a globally optimal solution is found by linear fitting [40], [48]. In our case, for samples $n = w+1, \ldots N-h$, we are left with

$$\hat{\mathcal{S}} = \mathcal{B} \; \Phi, \tag{7}$$

where $\mathcal{B} = [\mathbf{b}_0 \mathbf{b}_1 \ldots \mathbf{b}_K]$ is a $D \times (K+1)$ coefficient matrix and $\Phi = [\Phi_{w+1} \ldots \Phi_{N-h}]$ is formed by column vectors $\Phi_n = \left(1, \phi\left(\|\mathbf{s}_n^* - \mathbf{c}_1\|\right), \ldots \phi\left(\|\mathbf{s}_n^* - \mathbf{c}_K\|\right)\right)^{\mathrm{T}}$.

4) Locally constant: A zeroth-order approximation to the time series is given by a locally constant predictor [49]. With this predictor, one first determines a neighborhood Ω_n of radius ϵ around each point $\mathbf{s}_n^*$ of the reconstructed state space. Then forecasts

$$\hat{\mathbf{s}}_{n+h} = \frac{1}{|\Omega_n|} \sum_{\mathbf{s}_{n'} \in \Omega_n} \mathbf{s}_{n'+h}, \tag{8}$$

where $|\Omega_n|$ denotes the number of elements in Ω_n. In our prediction trials, ϵ is set to a percentage ϵ_κ of the mean distance between all state space points (we use the squared Euclidean norm). In addition, we require $|\Omega_n| \geq \nu$, i.e. a minimum of ν neighbors is always included independently of their distance to $\mathbf{s}_n^*$. Notice that this is almost a model-free approach with no coefficients to be learned: one just needs to set parameters m, τ, ϵ_κ and ν.

5) Naïve Markov: This approach is based on grouping inputs $\mathcal{S}^*$ and outputs $\hat{\mathcal{S}}$ into K_i and K_o clusters, respectively [41]. Given this partition, we fill in a $K_\mathrm{i} \times K_\mathrm{o}$ transition matrix $\mathcal{P}$, whose elements $p_{k_\mathrm{i},k_\mathrm{o}}$ correspond to the probability of going from cluster k_i of $\mathcal{S}^*$ to cluster k_o of $\hat{\mathcal{S}}$ (i.e. the rows of $\mathcal{P}$ sum up to 1). Then, when forecasting, a state space reconstruction $\mathbf{s}_n^*$ of the input $\mathbf{s}_n$ is formed and the distance towards all K_i input cluster medoids is calculated. In order to evaluate the performance of the Markov predictor in the same way as the other predictors, we use $\mathcal{P}$ to construct a deterministic output in the following way:

$$\hat{\mathbf{s}}_{n+h} = \sum_{k_\mathrm{o}=1}^{K_\mathrm{o}} p_{k_\mathrm{i}',k_\mathrm{o}}\ \mathbf{c}_{k_\mathrm{o}}, \tag{9}$$

where $\mathbf{c}_{k_\mathrm{o}}$ denotes the medoid of (output) cluster k_o and k_i' is the index of the (input) cluster whose medoid is closest to $\mathbf{s}_n^*$.

D. Training and testing

All previous models are completely described by a series of parameters (m, τ, K, α, ϵ_κ, ν, K_i or K_o) and coefficients ($\mathcal{A}$, $\mathcal{A}^{(k)}$, $\mathcal{B}$, $\mathcal{P}$, $\mathbf{c}_k$ or ρ_k). In our prediction trials these values are learned independently for each song *and* descriptor using the entire time series as training set. This learning is done with no prior information about parameters and coefficients. More specifically, for each song and descriptor time series we calculate the corresponding model coefficients for different parameter configurations and then select the solution that leads to the best in-sample approximation of the data. We perform a grid search over $m \in [1,\ 2,\ 3,\ 5,\ 7,\ 9,\ 12,\ 15]$ and $\tau \in [1,\ 2,\ 6,\ 9,\ 15]$ for all models, $K \in [1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 10,\ 12,\ 15,\ 20,\ 30,\ 40,\ 50]$ for TAR and RBF models, $\alpha \in [0.5,\ 0.75,\ 1,\ 1.25,\ 1.5,\ 2,\ 2.5,\ 3,\ 3.5,\ 4,\ 5,\ 7,\ 9]$ for RBF models, $\epsilon_\kappa \in [0.01,\ 0.025,\ 0.05,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7,\ 0.8]$ and $\nu \in [2,\ 5,\ 10,\ 15,\ 25,\ 50]$ for the locally constant predictor and $K_\mathrm{i} \in [8,\ 15,\ 30,\ 40,\ 50,\ 60,\ 70]$ and $K_\mathrm{o} \in [5,\ 10,\ 20,\ 30,\ 40,\ 50]$ for the naïve Markov method. Intuitively, with such a search for the best parameter combination for a specific song's time series, part of the dynamics' modeling is also done through the appropriate parameter setting.

Since we aim at obtaining compact descriptions of our data and we want to avoid overfitting, we limit the total number of model parameters and coefficients to be less than 10% of the total number of values of the time series data. This implies that parameter combinations leading to models with more than $(N \times D)/10$ values are automatically discarded at the training phase[3]. We also force an embedding window $w < N/20$.

[3]Of course this does not apply for the locally constant predictor, which, as already said, is a quasi model-free approach.

III. EVALUATION

A. Music data

We use an in-house music collection consisting of 2125 commercial songs (real-world recordings). In particular, we use an arbitrarily selected but representative compilation of *cover* songs. This music collection is an extension of the one used in our previous work [20] and it includes 523 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 4.06, ranging from 2 to 18. The collection spans a variety of genres, with their corresponding sub-genres and styles: pop/rock (1226 songs), electronic (209), jazz/blues (196), world music (165), classical music (133) and miscellaneous (196). Songs have an average length of 3.6 min, ranging from 0.5 to 8 min. For time-consuming parts of the analysis, a randomly selected subset of 102 songs is used (17 cover sets of cardinality 6, the same subset is used in all experiments).

B. Prediction error

To evaluate the predictability of the considered time series we use a normalized mean squared error measure [40], both when training our models (to select the best parameter combination) and when forecasting. We employ

$$\xi = \frac{1}{N-h-w} \sum_{n=w+1}^{N-h} \frac{1}{D} \sum_{d=1}^{D} \frac{\left(\hat{s}_{d,n+h} - s_{d,n+h}\right)^2}{\sigma_d^2}, \tag{10}$$

where σ_d^2 is the variance of the d-th descriptor component over all samples $n = w + h + 1, \ldots N$ of the target time series $\mathcal{S}$. Eq. (10) is a common way to measure the goodness of a prediction in the time series literature [34], [39], [40], [49]. Under the assumption of Gaussian errors which are independent of each other and of the data, the minimization of the mean squared error is mathematically equivalent to the maximization of the likelihood that a given model has generated the observed data.

We use the notation $\xi_{i,i}$ when a model trained on song i is used to forecast further frames of song i (self-prediction, in-sample error) and $\xi_{i,j}$ when a model trained on song i is used to forecast frames of song j (cross-prediction, out-of-sample error). In the case of self-prediction, we report the average error across songs, which we denote as $\langle \xi \rangle$. In the case of cross-prediction, each $\xi_{i,j}$ across all musical pieces is used to obtain an accuracy measure (see below).

C. Cover song identification

To evaluate the accuracy in identifying cover songs we proceed as in our previous work [20]. Given a music collection with I songs, we calculate $\xi_{i,j}$ for all $I \times I$ possible pairwise combinations and then create

a symmetric dissimilarity matrix $\mathcal{D}$, whose elements are $d_{i,j} = \xi_{i,j} + \xi_{j,i}$. Once $\mathcal{D}$ is computed, we resort to standard information retrieval (IR) measures to evaluate the discriminative power of this information. We use the *mean of average precisions* measure [50], which we denote as $\langle\overline{\psi}\rangle$. This measure is routinely employed in the IR [50] and MIR [14] communities and, in particular, in the cover song identification task [16].

To calculate $\langle\overline{\psi}\rangle$, $\mathcal{D}$ is used to compute a list $\mathit{\Lambda}_i$ of $I-1$ songs sorted in ascending order with regard to their dissimilarity to the query song i. Suppose that the query song i belongs to a cover set comprising $C_i + 1$ songs. Then, the average precision $\overline{\psi}_i$ is obtained as

$$\overline{\psi}_i = \frac{1}{C_i} \sum_{r=1}^{I-1} \psi_i(r) \mathit{\Gamma}_i(r), \tag{11}$$

where $\psi_i(r)$ is the precision of the sorted list $\mathit{\Lambda}_i$ at rank r,

$$\psi_i(r) = \frac{1}{r} \sum_{u=1}^{r} \mathit{\Gamma}_i(u), \tag{12}$$

and $\mathit{\Gamma}_i$ is the so-called relevance function: $\mathit{\Gamma}_i(v) = 1$ if the song with rank v in the sorted list is a cover of i and $\mathit{\Gamma}_i(v) = 0$ otherwise. Hence $\overline{\psi}_i$ ranges between 0 and 1. If the C_i covers of song i take the first C_i ranks, we get $\overline{\psi}_i = 1$. If all cover songs are found towards the end of $\mathit{\Lambda}_i$, we get $\overline{\psi}_i \approx 0$. The mean of average precisions $\langle\overline{\psi}\rangle$ is calculated as the mean of $\overline{\psi}_i$ across all queries $i = 1, \ldots I$. Using Eqs. (11) and (12) 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 dissimilarity matrix $\mathcal{D}$ has no discriminative power with regard to the assignment of cover songs. For this purpose, we separately permute $\mathit{\Lambda}_i$ for all i and keep all other steps the same. We repeat this process 99 times, corresponding to a significance level of 0.01 of this Monte Carlo null hypothesis test [51], and take the average, resulting in $\langle\overline{\psi}\rangle_{\text{null}}$. This $\langle\overline{\psi}\rangle_{\text{null}}$ is used to estimate the accuracy of all considered models under the specified null hypothesis.

D. Baseline predictors

Besides models in Sec. II-C, we further assess our results with a set of baseline approaches that do not require parameter adjustments nor coefficient determination.

1) Mean: The prediction is simply the mean of the training data:

$$\hat{\mathbf{s}}_{n+h} = \boldsymbol{\mu}, \tag{13}$$

$\boldsymbol{\mu}$ being a column vector. This predictor is optimal in the sense of Eq. (10) for i.i.d. time series data. Notice that, by definition, $\xi = 1$ when predicting with the mean of the time series data. In fact, ξ allows to estimate, in a variance percentage, how our predictor compares to the baseline prediction given by Eq. (13).

2) Persistence: The prediction corresponds to the current value:

$$\hat{\mathbf{s}}_{n+h} = \mathbf{s}_n. \tag{14}$$

This prediction yields low ξ values for processes that have strong correlations at h time steps.

3) Linear trend: The prediction is formed by a linear trend based on the current and the previous samples:

$$\hat{\mathbf{s}}_{n+h} = 2\mathbf{s}_n - \mathbf{s}_{n-1}. \tag{15}$$

This is suitable for a smooth signal and a short prediction horizon h.

IV. Results

A. Self-prediction

We first look at the average self-prediction error $\langle\xi\rangle$ one step ahead of the current sample, i.e. at horizon $h = 1$, corresponding to 116 ms (Table I). We see that, for all considered models, we do not achieve a drastic error reduction compared to the mean predictor (for which $\xi = 1$, Sec. III-D). However, in the majority of cases, all models are considerably better than the baselines. In particular, average errors $\langle\xi\rangle$ below 0.5 are achieved by the RBF, AR and TAR models. The latter is found to be the best forecast model across all descriptors. The fact that the predictability is weak but still better than the baseline provides evidence that music descriptors possess dependencies which can be exploited by deterministic models.

Remarkably, the above fact is observed independently for all models, musical facets and descriptors (Table I). Nevertheless, RT and OC descriptors have a considerably low $\langle\xi\rangle$ compared to the rest. This is due to the way these descriptors are computed. RT and OC are rhythm descriptors, and a characterization of such a musical facet cannot be captured in, say, 100 or even 500 ms. Humans need a longer time span to conceptualize rhythm [31] and, consequently, general rhythm descriptors use a longer analysis window. In particular we use for RT and OC a window of 6 s (Sec. II-B). Since we use a fixed step size to 116 ms, we have much stronger correlations. Indeed, the very low error we obtain already with the persistence and the linear trend predictors illustrates this fact. In addition, the genre configuration of

TABLE I

AVERAGE SELF-PREDICTION ERROR $\langle \xi \rangle$ FOR $h = 1$ WITH ALL DESCRIPTORS CONSIDERED (FULL MUSIC COLLECTION). THE MEAN PREDICTOR IS NOT SHOWN SINCE, BY DEFINITION, ITS ERROR EQUALS 1 (SEC. III-D).

Methods	Descriptors											
	PCP	TC	HC	MFCC	SC	SP	SH	OP	RT	OC	E	L
Linear trend	2.334	1.869	1.380	1.816	1.444	1.878	3.336	2.079	0.012	0.103	1.839	2.148
Persistence	0.861	0.718	0.729	0.706	0.550	0.683	1.191	0.747	0.013	0.070	0.675	0.743
Naïve Markov	0.770	0.660	0.534	0.804	0.654	0.455	0.819	0.498	0.633	0.757	0.474	0.616
Locally constant	0.647	0.574	0.804	0.637	0.564	0.799	0.418	0.452	0.334	0.587	0.460	0.423
RBF	0.575	0.502	0.416	0.492	0.343	0.380	0.729	0.429	0.030	0.124	0.409	0.330
AR	0.527	0.480	0.444	0.416	0.290	0.384	0.759	0.438	0.002	0.014	0.421	0.339
TAR	**0.480**	**0.439**	**0.387**	**0.415**	**0.276**	**0.341**	**0.697**	**0.390**	**0.001**	**0.010**	**0.376**	**0.306**

our music collection might explain part of this low error: more than $2/3$ of our collection is classified between the pop/rock and electronic genres, which are characterized by more or less plain rhythms.

We now study the average self-prediction error $\langle \xi \rangle$ as the forecast horizon h increases (Fig. 2). We see that $\langle \xi \rangle$ increases rapidly for $h \leq 4$ (or 10, depending on the descriptor) but, surprisingly, it reaches a stable plateau with all descriptors for $h > 10$, i.e. for prediction horizons of more than 1 s. Notably, in this plateau, $\langle \xi \rangle < 1$. This indicates that, on average, there is a certain capability for the models to still perform predictions at relatively long horizons, and that these predictions are better than predicting with the mean. This reveals that descriptor time series are far from being i.i.d. data (even at relatively long h) and that models are capturing part of the long-term structures and repetitions found in our collection's songs. If we analyze different musical facets, we can roughly say that rhythm descriptors are better predictable than the rest. Then come loudness descriptors and afterwards the timbre ones, although MFCC and SH descriptors are in the same error range as tonal descriptors. Indeed, tonal descriptors seem to be more difficult to predict. Still, we see that $\langle \xi \rangle < 0.8$ for PCP and TC descriptors for $h \leq 30$.

Overall, Fig. 2 evidences that music has a certain predictability. In particular, it reflects this characteristic for a broad range of prediction intervals. All this is confirmed independently of the model, the descriptor and therefore of the musical facet considered. In addition, if we pay attention to the behavior of all curves in Fig. 2, we see that the error grows sub-linearly (the curves resemble the ones for a square root or a logarithm). This can be observed for all models and descriptors tested. Further information on the

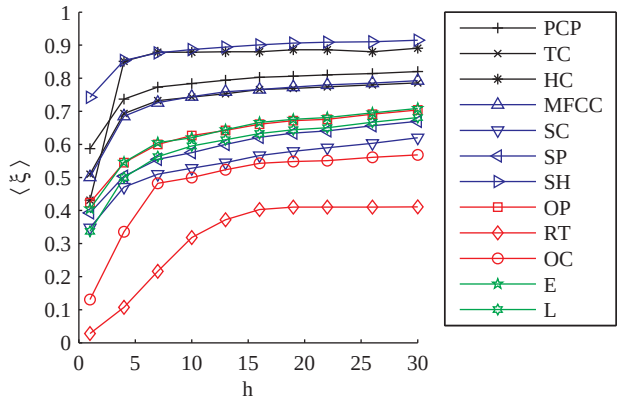

Fig. 2. Average self-prediction error $\langle \xi \rangle$ as a function of the prediction horizon h. This is obtained with the RBF method for all considered descriptors (102-song collection). Other methods yield qualitatively similar plots.

behavior of these curves is provided as Supplementary Material[4], where we also derive some conjectures regarding the sources of predictability of music descriptor time series. In the Supplementary Material we also suggest that the behavior of these time series may be reproduced by a concatenation of multiple AR processes with superimposed noise.

Next, we discuss about the best parameter combinations for each model and descriptor; in particular for embedding values and number of clusters. In general, few clear tendencies could be inferred. One is that AR and TAR models select as optimal relatively high values for m and τ with nearly all descriptors. Other combinations of descriptors and models tend to use intermediate or lower values among the ones tested (Sec. II-C). In particular, the naïve Markov and the locally constant predictors tend to use m and τ values both between 1 and 3. The number of clusters K for RBF and TAR models (or K_{i} and K_{o} for the naïve Markov approach) practically always reaches the imposed limitation that a model has to be described by a maximum of $1/10$ of the original data.

If we check for individual song's predictability, we can see some curious examples. For instance, many renditions of the theme "Radioactivity", originally performed by Kraftwerk, and "Little 15", originally performed by Depeche Mode, achieve quite small ξ values with the TC descriptor, i.e. their tonal behavior seems somewhat predictable. Indeed, these musical pieces possess a slow tempo, highly repetitive, simple tonal structures and can be classified into the pop/electronic genres. On the other hand, high ξ values for the TC descriptor are encountered in many jazz covers (with relatively long improvisations) or in some

[4]http://mtg.upf.edu/files/personal/IEEEPredictabilitySupplementary.pdf

versions of "Helter skelter", originally performed by The Beatles, for which we have some heavy-metal cover songs[5] in our collection with quite up-tempo long guitar solos. Analogous observations can be made for timbre and rhythm descriptors. For example, several renditions of "Ave Maria" (both Bach's and Schubert's compositions) performed by string quartets or similar instrumental formations lead to low ξ values with the MFCC descriptor (this indicates that timbres do no change too much within the piece). With regard to rhythm, we find performances of "Bohemian rhapsody", originally performed by Queen, to yield relatively high ξ values with the RT and OC descriptors (there are clearly marked rhythm changes in the composition).

B. Cross-prediction

To assess whether a model-based cross-prediction strategy is useful for cover song detection we study the mean of average precisions $\langle\overline{\psi}\rangle$ obtained from $\mathcal{D}$, the symmetrized version of the cross-prediction errors (Sec. III-C). As before, we consider different prediction horizons h. Since tonality is the main musical facet exploited by cover song identification systems (Secs. I and II-A), in this section we just consider PCP, TC and HC descriptors.

In Fig. 3 we see that, except for the locally constant predictor, all models perform worse than the mean predictor for short horizons ($h \leq 3$). This performance increases with the horizon ($4 \leq h \leq 7$), but reaches a stable value for mid-term and relatively long horizons ($h > 7$), which is much higher than the mean predictor performance. Remarkably, in previous Sec. IV-A we show that for $h > 7$, PCP and TC descriptors yield an average prediction error $\langle\xi\rangle < 0.8$, which denotes the capability of all considered models to still perform predictions at relatively long horizons. We now assert that this fact, which to the best of the authors' knowledge has never been reported in the literature, becomes crucial for cover song identification (Fig. 3).

The fact that we better detect cover songs at mid-term and relatively long horizons could possibly have a musicological explanation. To see this we study matrices quantifying the transition probabilities between states separated by a time interval corresponding to the prediction horizon h. We first cluster a time series $\mathcal{S}$ into, say, 10 clusters and compute the medoids. We subsequently fill a transition matrix $\mathcal{T}$, with elements $t_{i,j}$. Here i and j correspond to the indices of the medoids to which respectively $\mathbf{s}_n$ and $\mathbf{s}_{n+h}$ are closest. This transition matrix is normalized so that each row sums to 1. In Fig. 4 we show $\mathcal{T}$ for three different horizons ($h = 1$ in the first column, $h = 7$ in the second column and $h = 15$ in

[5]Actually "Helter skelter" could be considered one of the pioneering songs of heavy-metal.

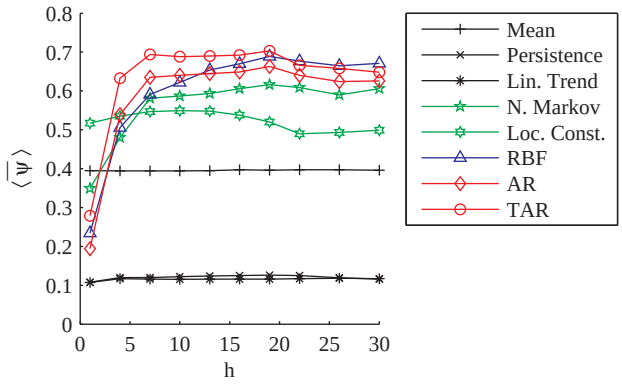

Fig. 3. Mean of average precisions $\langle\overline{\psi}\rangle$ in dependence on the prediction horizon h. Results for the TC descriptor with all considered models (102-song collection). PCP and HC time series yield qualitatively similar plots.

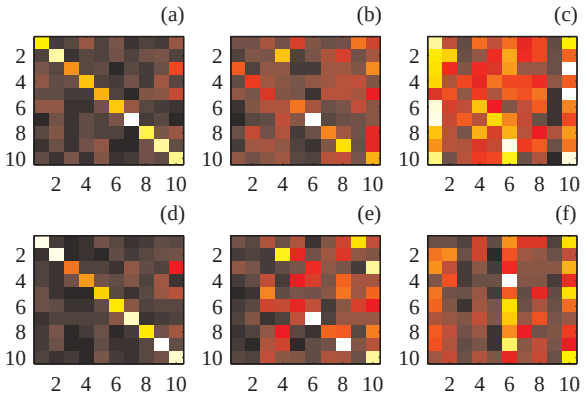

Fig. 4. Transition matrices $\mathcal{T}$ for two cover songs (top, one song per axis) and two unrelated songs (bottom). These are computed for $h = 1$ (a,d), $h = 7$ (b,e) and $h = 15$ (c,f). Bright colors correspond to high transition probabilities (white and yellow patches).

the third column). Two unrelated songs are shown (one row each). The musical piece that provided the cluster medoids to generate $\mathcal{T}$ is a cover of the first song (top row) but not of the second one (bottom row).

With this small test we see that, for $h = 1$, $\mathcal{T}$ is highly dominated by persistence to the same cluster, both for the cover (Fig. 4a) and the non-cover (Fig. 4d) pair. This fact is also seen with the self-prediction results of the persistence-based predictor (Table I). Once h increases, characteristic transition patterns arise, but the similarity between matrices in Fig. 4b and 4e shows that these patterns are not characteristic enough to define a song. Compare for example the high values obtained for both songs in $t_{7,6}$, $t_{9,8}$, $t_{2,4}$,

TABLE II

MEAN OF AVERAGE PRECISIONS $\langle\overline{\psi}\rangle$ FOR THE COVER SONG IDENTIFICATION TASK (FULL MUSIC COLLECTION). THE MAXIMUM OF THE RANDOM BASELINE $\langle\overline{\psi}\rangle_{\text{NULL}}$ WAS FOUND TO BE 0.008 WITHIN 99 RUNS.

Methods	Descriptors		
	PCP	TC	HC
Linear trend	0.007	0.007	0.006
Persistence	0.006	0.007	0.006
Mean	0.146	0.094	0.013
Locally constant	0.254	0.280	0.046
Naïve Markov	0.367	0.379	0.049
AR	0.368	0.407	0.044
RBF	0.377	0.438	0.054
TAR	**0.386**	**0.441**	**0.064**

$t_{1,9}$ or $t_{3,10}$. We conjecture that these transitions define general musical features that are shared among a big subset of songs, not necessarily just the covers. For example, it is clear that there are general rules with regard to chord transitions, with some particular transitions being more likely than others [3], [4]. Only when $h > 7$ transitions that discriminate between the dynamics of songs start to become (see the distinct patterns in Figs. 4c and 4f). This distinctiveness can then be exploited to differentiate between cover and non-cover songs.

Results in detecting cover songs with the full collection (Table II) indicate that the best model is, as with the self-prediction trials, the TAR model; although notable accuracies are achieved with the RBF method. The AR and the naïve Markov models come next. Persistence and linear trend predictors perform at the level of the random baseline $\langle\overline{\psi}\rangle_{\text{null}}$. This is to be expected since no learning is performed for these predictors. In addition, we see that the HC descriptor is much less powerful than the other two. This is again to be expected, since HC compresses tonal information to a univariate value. Furthermore, HC might be less informative than PCP or TC values themselves, which already contain the change information in their temporal evolution. Apart from this, we see that TC descriptors perform better than PCP descriptors. This does not necessarily imply that TC descriptors provide a better representation of a song's tonal information, but that TAR models are better in capturing the essence of their temporal evolution.

V. Discussion: model-based cover song detection

Even though the considered models yield a significant accuracy increase when compared to the baselines, it might still seem that a value of $\langle\overline{\psi}\rangle$ around 0.4 in an evaluation measure that ranges between 0 and 1 is not a big success for a cover song identification approach. To properly asses this accuracy one has to compare it against the accuracies of state-of-the-art approaches. According to an international MIR evaluation framework (the yearly MIR evaluation exchange, MIREX [14]), the best accuracy achieved to date within the cover song identification task[6] was obtained from a previous system by Serrà et al. [20]. This system reached $\langle\overline{\psi}\rangle = 0.66$ with the MIREX dataset and yields $\langle\overline{\psi}\rangle = 0.698$ with the music collection used here. A former method by Serrà et al. [19] scored $\langle\overline{\psi}\rangle = 0.55$ with the MIREX data. Thus the cross-prediction approach does not outperform these methods. However, cited methods were specifically designed for the task of identifying cover songs, while the cross-prediction approach is a general schema that does not incorporate specific modifications that could be beneficial for such a task [16] (e.g. taking into account tempo or structural changes between cover songs). To make further comparisons (at least qualitatively), one should note that $\langle\overline{\psi}\rangle$ values around 0.4 are in line with other state-of-the-art accuracies, or even better if we consider comparable music collections [16].

Beyond accuracy comparisons, some other aspects can be discussed. Indeed, another reason for appraising the solution obtained here comes from the consideration of storage capabilities and computational complexities at the query retrieval stage. Since we limit our models to a size of 10% of the total number of training data (Sec. II-C), they require 10% of the storage that would be needed for saving the entire time series (state-of-the-art systems usually store the full time series for each song). This fact could be exploited in a single-query retrieval scenario. In this setting, it would be sufficient to determine a dissimilarity measure ξ (Eq. 10) from the application of all models to the query song. Hence, only the models rather than the raw data would be required. Regarding computational complexity, many approaches for cover song identification are quadratic in the length of the time series, requiring at least a Euclidean distance calculation for every pair of sample points [16] (e.g. [19], [20]). In contrast, the approaches presented here are linear in the length of the time series. For example, with TAR models, we just need to do a pairwise distance calculation between the samples and the K medoids, plus a matrix multiplication and subtraction (notice that the former is not needed with AR models). If we compare the previous approach [20] with the TAR-based strategy by considering an average time series length $\bar{N}$, we have that

[6]These correspond to the 2008 and 2009 editions, which are available from `http://music-ir.org/mirex/2008` and `http://music-ir.org/mirex/2009`, respectively.

the former is roughly $O\left(\bar{N}^2 mD\right)$, while the latter is $O\left(\bar{N} mD(K+D)\right)$, with $K + D \ll \bar{N}$. To put some numbers: with $\bar{N} = 2304$ (approximately 4 min of music), descriptor dimensionality $D = 12$ (the largest among PCP, TC and HC, Sec. II-C) and $K = 50$ (the maximum allowed), we obtain a minimal relative speed improvement of $2304/(50 + 12) \approx 37$.

A further and very interesting advantage of using the approaches considered here is that no parameters need to be adjusted by the user. More specifically, models' parameters and coefficients are automatically learned for each song and descriptor time series individually by the minimization of the in-sample training error $\xi_{i,i}$. Usually, cover song identification algorithms have multiple parameters that can be dependent, for instance, on the music collection, the music descriptor time series or the types of cover songs under consideration [16]. Previously cited methods [19], [20] were not an exception: as there was no way to a priori set their specific parameters, these were set by trial and error with an independent out-of-sample music collection. Since for the current approaches no such manual parameter optimization is required, its application is robust and straightforward.

In conclusion, we see that considering cross-predictions of music descriptor time series leads to parameter-free approaches for cover song identification that are substantially faster, allow for reduced computational storage and still maintain highly competitive accuracies when compared to state-of-the-art systems. Thus, the use of the concept of cross-prediction errors stands as a promising strategy for cover song detection and, by extension, for music retrieval in general.

VI. Summary and future perspectives

In the present work we take an interdisciplinary approach to study music predictability, encompassing concepts and methods from signal processing, music technology, linear and nonlinear time series modeling, machine learning and information retrieval. We first apply current signal processing and music technology approaches for extracting meaningful information from real audio signals in the form of music descriptor time series. We then explore a number of modeling strategies to assess the predictability of these time series. We test a number of routinely employed time series models. These comprise linear and nonlinear predictors, namely AR, TAR, RBF, locally constant and naïve Markov. These models are automatically trained for each song and descriptor time series individually. Training is done by performing a grid search over a set of parameters and automatically determining the corresponding coefficients.

First, we perform an in-sample self-prediction of descriptor time series extracted from a representative music collection. This allows us to assess which modeling strategy gives a lower prediction error. TAR, AR and RBF methods provide the best predictions. Furthermore, our analysis allows us to quantitatively assess

the predictability of different musical facets represented by the descriptors. Overall, rhythm descriptors (RT and OC) are better predictable than the rest. All tonal descriptors (PCP, TC and HC), together with some timbral ones (MFCC and SH) are more difficult to predict.

Some general features are common to all considered descriptors, musical facets and models. The prediction error behaves sub-linearly, resembling a square-root or logarithmic curve when plotted against the prediction horizon. Despite being relatively high in absolute values, the prediction error is still lower than the one expected for random data and unspecific predictors. This is observed for a short as well as mid-term and relatively long prediction horizons. To the best of the authors' knowledge, these aspects have not been assessed before.

These findings provide quantitative evidence for the qualitative, intuitive notion that music has both predictable and unpredictable components. It is important, however, to emphasize that our work does not represent a formal testing of some null hypothesis about the process underlying music signals. For this purpose we would at first need to specify a stochastic model for this process. For example, we could test the null hypothesis that music descriptor time series are consistent with the output of a linear stochastic Gaussian process. We would need to specify all assumptions made about this process, for example with regard to its autocorrelation, cross correlation or stationarity. To test this null hypothesis we would then need to generate surrogate data from our original descriptor time series [52]. These surrogates would be constructed to be identical to the original time series with regard to properties that are included in the null hypothesis (e.g. surrogates can be constructed to have the same autocorrelation and cross correlation as the original data but to be otherwise random). Finally, we would need to extract some discriminating statistics such as a nonlinear prediction error from the original time series and the surrogates [34]. If the results obtained for the original time series were significantly different from the ones of the surrogates, the null hypothesis could be rejected.

However, such a formal null hypothesis testing is not the aim of the present study. Rather, we provide a real-world application of out-of-sample cross-prediction errors. We do so by addressing the information retrieval task of cover song identification In particular, we see that reliability of models' predictions at mid-term and relatively long horizons allows us to effectively perform this task. Moreover, we show that a model-based cross-prediction strategy for cover song identification achieves competitive accuracies when compared to the state-of-the-art, while requiring less computational time and storage (e.g. TAR, RBF and AR models). Importantly, the proposed approach is parameter-free from a user's perspective, since all necessary parameters are automatically set for each song and descriptor time series individually. This makes the application of the proposed approaches robust and straightforward.

Future lines of research include further investigation on model-based cross-prediction strategies for cover song detection. Given the benefits of model-based approaches discussed in Sec. V, it is interesting to see whether further refinements could yield accuracies that surpass the ones achieved by previous work [19] and [20]. In particular, it would be interesting to see how hidden Markov models (HMM) [53] can be adapted to cover song identification. Such models have been very successfully applied within the speech processing community [21] and also to music processing tasks (e.g. in tempo and pitch tracking [12]). However, while descriptor time series have values on continuous scales, HMMs operate on discrete sets of states. Since the discretization of descriptor time series is not straightforward in the case of cover songs [16], one may consider a continuous version of such models. In addition, we conjecture that specific adaptations are needed, in particular adaptations dealing with tempo and structure invariance (see e.g. [54] for some structure invariance adaptations of HMMs in the related context of exact audio matching). Indeed, these are two important characteristics that a cover song identification system has to take into account [16]. With regard to tempo invariance, we hypothesize that working with tempo-insensitive representations of tonal information (e.g. [17]) can partially solve the problem. However, one should take care of the beat-detection stage used to obtain such representations, since it might introduce additional errors to the system [16]. Notice that the introduction of a tempo invariant representation is only one option and that further strategies can be devised, specially with the setting of the prediction horizon h and the time delay τ. With regard to structure invariance, the easy way would be to cut the time series into short overlapping segments (e.g. [18]) and train different models on each segment. However, this solution would introduce additional computational costs since each error for each segment would need to be evaluated. Last but not least, one could investigate with the incorporation of musical knowledge in order to yield better predictions.

ACKNOWLEDGMENTS

The authors would like to thank Jochen Bröker for useful discussions, Perfecto Herrera for useful discussions, comments and proofreading and Justin Salamon for comments and proofreading.

REFERENCES

[1] S. Mithen, *Singing Neanderthals: the Origins of Music, Language, Mind and Body, The*. Cambridge, USA: Harvard University Press, 2007.

[2] D. J. Levitin, “Why music moves us,” *Nature*, vol. 464, no. 8, pp. 834–835, 2010.

[3] D. Huron, *Sweet anticipation: music and the psychology of expectation*. Cambridge, USA: MIT Press, 2006.

[4] E. Narmour, *Analysis and Cognition of Basic Melodic Structures: the Implication-Realization Model, The*. Chicago, USA: University of Chicago Press, 1990.

[5] F. Pachet, "The continuator: musical interaction with style," *Journal of New Music Research*, vol. 31, no. 1, pp. 1–9, 2002.

[6] T. Eerola, P. Toiviainen, and C. L. Krumhansl, "Real-time prediction of melodies: continuous predictability judgements and dynamic models," *Int. Conf. on Music Perception and Cognition (ICMPC)*, pp. 473–476, 2002.

[7] S. Abdallah and M. D. Plumbey, "Information dynamics: patterns of expectation and surprise in the perception of music," *Connection Science*, vol. 21, no. 2, pp. 89–117, 2009.

[8] J. F. Paiement, Y. Grandvalet, and S. Benigo, "Predictive models for music," *Connection Science*, vol. 21, no. 2, pp. 253–272, 2009.

[9] S. Dubnov, "Spectral anticipations," *Computer Music Journal*, vol. 30, no. 2, pp. 63–83, 2006.

[10] S. Dubnov, G. Assayag, and A. Cont, "Audio oracle: a new algorithm for fast learning of audio structures," *Int. Computer Music Conference (ICMC)*, pp. 224–228, 2007.

[11] A. Hazan, R. Marxer, P. Brossier, H. Purwins, P. Herrera, and X. Serra, "What/when causal expectation modelling applied to audio signals," *Connection Science*, vol. 21, no. 2, pp. 119–143, 2009.

[12] A. Klapuri and M. Davy, *Signal processing methods for music transcription*. New York, USA: Springer Science, 2006.

[13] N. Orio, "Music retrieval: a tutorial and review," *Foundations and Trends in Information Retrieval*, vol. 1, no. 1, pp. 1–90, 2006.

[14] J. S. Downie, "The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[15] M. Casey, R. C. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[16] J. Serrà, E. Gómez, and P. Herrera, "Audio cover song identification and similarity: background, approaches, evaluation and beyond," in *Adv. in Music Information Retrieval*, ser. Studies in Computational Intelligence, Z. W. Ras and A. A. Wieczorkowska, Eds. Berlin, Germany: Springer, 2010, vol. 16, no. 6, ch. 14, pp. 307–332.

[17] D. P. W. Ellis and G. E. Poliner, "Identifying cover songs with chroma features and dynamic programming beat tracking," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 2007, pp. 1429–1432.

[18] M. Marolt, "A mid-level representation for melody-based retrieval in audio collections," *IEEE Trans. on Multimedia*, vol. 10, no. 8, pp. 1617–1625, 2008.

[19] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1138–1152, 2008.

[20] J. Serrà, X. Serra, and R. G. Andrzejak, "Cross recurrence quantification for cover song identification," *New Journal of Physics*, vol. 11, p. 093017, 2009.

[21] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*. Upper Saddle River, USA: Prentice Hall, 1993.

[22] J. Serrà, E. Gómez, and P. Herrera, "Transposing chroma representations to a common key," in *Proc. of the IEEE CS Conf. on The Use of Symbols to Represent Music and Multimedia Objects*, 2008, pp. 45–48.

[23] E. Gómez, P. Herrera, P. Cano, J. Janer, J. Serrà, J. Bonada, S. El-Hajj, T. Aussenac, and G. Holmberg, "Music similarity systems and methods using descriptors," *Patent US 2008/0300702*, December 4 2008.

[24] E. Gómez, "Tonal description of music audio signals," Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona, Spain, 2006, available online: http://mtg.upf.edu/node/472.

[25] C. Harte, M. B. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proc. of the ACM Workshop on Audio and Music Computing Multimedia*, 2006, pp. 21–26.

[26] M. Slaney, "Auditory toolbox. version 2," Tech. Rep., 1998, available online: http://cobweb.ecn.purdue.edu/malcolm/interval/1998-010.

[27] V. Akkermans, J. Serrà, and P. Herrera, "Shape-based spectral contrast descriptor," in *Proc. of the Sound and Music Computing Conf. (SMC)*, 2009, pp. 143–148.

[28] P. Brossier, "Automatic annotation of musical audio for interactive applications," Ph.D. dissertation, Queen Mary, University of London, London, UK, 2007, available online: http://aubio.org/phd.

[29] E. Guaus and P. Herrera, "The rhythm transform: towards a generic rhythm description," in *Proc. of the Int. Computer Music Conference (ICMC)*, no. 1131, 2005.

[30] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer, "On rhythm and general music similarity," in *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2009, pp. 525–530.

[31] P. Desain and H. Honing, "The formation of rhythmic categories and metric priming," *Perception*, vol. 32, no. 3, pp. 341–365, 2003.

[32] A. V. Oppenheim, R. W. Schafer, and J. B. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, USA: Prentice Hall, 1999.

[33] E. Vickers, "Automatic long-term loudness and dynamics matching," in *Proc. of the Conv. of the Audio Engineering Society (AES)*, no. 5495, 2001.

[34] H. Kantz and T. Schreiber, *Nonlinear time series analysis*, 2nd ed. Cambridge, UK: Cambridge University Press, 2004.

[35] F. Takens, "Detecting strange attractors in turbulence," *Lecture Notes in Mathematics*, vol. 898, pp. 366–381, 1981.

[36] I. Mierswa and K. Morik, "Automatic feature extraction for classifying audio data," *Machine Learning Journal*, vol. 58, pp. 127–149, 2005.

[37] F. Mörchen, A. Ultsch, M. Thies, and I. Löhken, "Modelling timbre distance with temporal statistics from polyphonic music," *IEEE Trans. on Speech and Audio Processing*, vol. 14, no. 1, pp. 81–90, 2006.

[38] G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control*, rev. ed. Oakland, USA: Holden-Day, 1976.

[39] H. Lütkepohl, *New introduction to multiple time series analysis*. Berlin, Germany: Springer, 2005.

[40] A. S. Weigend and N. A. Gershenfeld, *Time Series Prediction: Forecasting the Future and Understanding the Past*. Boulder, USA: Westwiew Press, 1993.

[41] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd ed. Amsterdam, The Netherlands: Elsevier, 2007.

[42] H. S. Parka and C. S. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[43] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 3rd ed. San Diego, USA: Academic Press, 2006.

[44] W. H. Press, B. P. Flannery, S. A. Tenkolsky, and W. T. Vetterling, *Numerical Recipes*, 2nd ed. Cambridge, UK: Cambridge University Press, 1992.

[45] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, "Temporal feature integration for music genre classification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1654–1664, 2007.

[46] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 174–186, 2009.

[47] H. Tong and K. S. Lim, "Threshold autoregression, limit cycles and cyclical data," *Journal of the Royal Statistical Society*, vol. 42, no. 3, pp. 245–292, 1980.

[48] D. S. Broomhead and D. Lowe, “Multivariable functional interpolation and adaptive networks,” *Complex Systems*, vol. 2, pp. 321–355, 1988.

[49] J. D. Farmer and J. J. Sidorowich, “Predicting chaotic time series,” *Physical Review Letters*, vol. 59, no. 8, pp. 845–848, 1987.

[50] C. D. Manning, R. Prabhakar, and H. Schutze, *An Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.

[51] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Berlin, Germany: Springer, 2004.

[52] T. Schreiber and A. Schmidt, “Surrogate time series,” *Physica D*, vol. 142, pp. 346–382, 2000.

[53] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*. New York, USA: Springer Science, 2005.

[54] E. Batlle, J. Masip, and E. Guaus, “Automatic song identification in noisy broadcast audio,” in *Proc. of the Signal and Image Processing Conf. (SIP)*, 2002, pp. 101–111.

Joan Serrà obtained both the degrees of Telecommunications and Electronics at Enginyeria La Salle, Universitat Ramón Llull, Barcelona, Spain, in 2002 and 2004, respectively. After working from 2005 to 2006 at the research and development department of Music Intelligence Solutions Inc, he joined the Music Technology Group of Universitat Pompeu Fabra, Barcelona, where he received the MSc and PhD in Information, Communication and Audiovisual Media Technologies in 2007 and 2011, respectively. He is currently a post-doc researcher with the Music Technology Group of the UPF. He is also a part-time associate professor with the Dept. of Information and Communication Technologies of the same university. In 2010 he was a guest scientist with the Research Group on Nonlinear Dynamics and Time Series Analysis of the Max Planck Institute for the Physics of Complex Systems in Dresden, Germany. His main research interests include music retrieval and understanding, signal processing, time series analysis, complex networks, complex systems, information retrieval, and music perception, psychology and cognition.

Holger Kantz is head of the Time Series Analysis group at the Max Planck Institute for the Physics of Complex Systems (MPIPKS) in Dresden, Germany. After studying Physics, he obtained his PhD in Theoretical Physics from Wuppertal University in 1989. Having spent several years as postdoctoral fellow, he joined the newly founded MPIPKS in 1995. He specialized in novel methods for the analysis of observed data, data classification, model identification and prediction. His scientific interests go beyond data analysis methods and include efforts to understand the sources of complex time dependencies in nonlinear stochastic processes and in systems with many degrees of freedom such as the atmosphere, with a strong emphasis on predictability and predictions. His reputation is based not only on a large number of research articles but also on his co-authorship of a well known textbook on time series analysis and the free software package TISEAN.

Xavier Serra is associate professor of the Department of Information and Communication Technologies and director of the Music Technology Group at the Universitat Pompeu Fabra in Barcelona. After a multidisciplinary academic education he obtained a PhD in Computer Music from Stanford University in 1989 with a dissertation on the spectral processing of musical sounds that is considered a key reference in the field. His research interests cover the understanding, modelling and generation of musical signals by computational means, with a balance between basic and applied research and approaches from both scientific/technological and humanistic/artistic disciplines. Dr. Serra is very active in promoting initiatives in the field of Sound and Music Computing at the local and international levels, being involved in the editorial board of a number of journals and conferences and giving lectures on current and future challenges of the field. He has recently been awarded an Advanced Grant of the European Research Council to carry out the project CompMusic aimed at promoting multicultural approaches in music computing research.

Ralph G. Andrzejak was born in Düsseldorf, Germany, in 1970. He graduated in Physics at the University of Bonn in 1997. He wrote his PhD at the Helmholtz-Institute for Radiation and Nuclear Physics and the Department of Epileptology, University of Bonn, Germany, in 2001. From 2002 to 2004 he carried out a first postdoc fellowship at the John-von-Neumann Institute for Computing, Research Center Jülich, Germany. For 2005 to 2006 he was awarded with a Feodor Lynen postdoc fellowship of the Alexander von Humboldt-Foundation through which he joined the Computational Neuroscience Group at the Department of Information and Communication Technologies of the Universitat Pompeu Fabra, Barcelona, Spain. For 2007 to 2011 he was awarded with a Ramón y Cajal fellowship of the Spanish Ministry for Science and Innovation which allowed him to continue his work at the Department of Information and Communication Technologies of the Universitat Pompeu Fabra. His main research interests are linear and nonlinear signal analysis techniques. An emphasis is placed on the characterization of directional couplings between dynamical systems. Apart from the study of signals from model systems he focuses on applications of signal analysis techniques to recordings from experimental dynamics.