Identifying Cover Songs Using Information-Theoretic Measures of Similarity

Peter Foster, *Student Member, IEEE*, Simon Dixon, and Anssi Klapuri

***Abstract*—This paper investigates methods for quantifying similarity between audio signals, specifically for the task of cover song detection. We consider an information-theoretic approach, where we compute pairwise measures of predictability between time series. We compare discrete-valued approaches operating on quantized audio features, to continuous-valued approaches. In the discrete case, we propose a method for computing the normalized compression distance, where we account for correlation between time series. In the continuous case, we propose to compute information-based measures of similarity as statistics of the prediction error between time series. We evaluate our methods on two cover song identification tasks using a data set comprised of 300 Jazz standards and using the Million Song Dataset. For both datasets, we observe that continuous-valued approaches outperform discrete-valued approaches. We consider approaches to estimating the normalized compression distance (NCD) based on string compression and prediction, where we observe that our proposed normalized compression distance with alignment (NCDA) improves average performance over NCD, for sequential compression algorithms. Finally, we demonstrate that continuous-valued distances may be combined to improve performance with respect to baseline approaches. Using a large-scale filter-and-refine approach, we demonstrate state-of-the-art performance for cover song identification using the Million Song Dataset.**

***Index Terms*—Audio similarity measures, cover song identification, normalized compression distance, time series prediction.**

I. Introduction

IN THE field of music content analysis, quantifying similarity between audio signals has received a substantial amount of interest [1]. Owing to the proliferation of music in digital formats, there exists potential for applications using music similarity techniques, in a wide range of domains. At the level of individual tracks, these domains span audio fingerprinting [2], cover song identification [3], artist identification [4], [5] and genre classification [6]. Applications can be distinguished according to their *degree of specificity* [1], referring to the level of granularity required for retrieving audio tracks from a collection, given a query track. For example, in audio fingerprinting, the required specificity is high, since the set of possible tracks corresponding to a particular recording is typically small, in relation to the data set. In contrast, genre classification requires low specificity, since the set of tracks sharing a common genre is potentially large, in relation to the data set.

A cover song may be defined as a rendition of a previously recorded piece of music [7]. Cover song identification is deemed to have mid-level, diffuse specificity, since cover songs may differ from the original song in various musical facets, including rhythm, tempo, melody, harmonization, instrumentation, lyrics and musical form. Correspondingly, cover song identification remains a challenging problem [3].

In this work, we investigate methods for cover song identification that are based on quantifying pairwise predictability between sequences. From a music-psychological perspective, the significance of intrinsic predictability in musical sequences has been reflected on by Meyer [8], who considers the possibility of using Shannon's information theory [9] to quantify predictive uncertainty. Statistical learning is implicated in forming musical expectations [10]; a successful approach to modelling expectations in response to an unfolding stream of musical events involves estimating sequential statistical models and computing information-theoretic measures of predictive uncertainty [11]. As exemplified in [12], an information-theoretic approach admits a rich conceptual framework for quantifying predictive uncertainty in musical sequences. For our own purposes in cover song identification, we seek to establish if an information-theoretic approach might be useful for determining pairwise similarity between tracks.

Based on our previous work [13], we consider an information-theoretic approach to quantifying similarity between feature vector sequences. One possible approach based on the non-Shannon information measure of Kolmogorov complexity [14], the normalized compression distance (NCD) [15], quantifies similarity between two strings in terms of joint compressibility. The NCD has been applied successfully across a range of problem domains [15]–[18], including music content analysis [19]–[24]. For our chosen task of cover song identification, we interpret the NCD as a measure of pairwise predictability. Using our information-theoretic framework, we compare the NCD to alternative predictability measures based on Shannon information. We provide an evaluation of competing information-theoretic approaches and identify issues concerning their implementation. This paper extends our previous work [13] as follows: Firstly, we examine a larger

Manuscript received June 08, 2014; revised November 24, 2014; accepted March 04, 2015. Date of publication March 25, 2015; date of current version April 10, 2015. The work of P. Foster was supported by an Engineering and Physical Sciences Research Council Doctoral Training Account studentship. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Emmanuel Vincent.

P. Foster and S. Dixon are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS U.K. (e-mail: peter.foster@eecs.qmul.ac.uk; simon.dixon@eecs.qmul.ac.uk).

A. Klapuri is with the Tampere University of Technology, Tampere, Finland, FI-33720 (e-mail: anssi.klapuri@tut.fi).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2416655

set of distance measures and estimate distance measures by predicting discrete-valued sequences. Further, we incorporate the Million Song dataset (MSD) [25] into our evaluations. Finally, we investigate combining distance measures using both our considered datasets.

The remainder of this paper is organized as follows: Section II discusses related work on audio-based cover song identification and methods for determining musical similarity. Section III introduces the pairwise similarity methods evaluated in this work. Section IV describes our experimental procedure. Finally, in Sections V and VI we present results and conclusions.

II. Related Work

A. Musical Similarity

Methods for characterizing similarity between sequences of audio features can be distinguished based on whether the temporal order of features is discarded or retained [1]. In the former so-called ‘bag-of-features’ approach, a widespread method involves estimating distributions of features obtained from time-frequency representations of musical audio [5], [26]–[31]. The bag-of-features approach is unable to model the temporal aspect of music, in which rhythmic, harmonic and melodic objects exhibit sequential structure and in which repetition and variation are of importance [32]. Casey and Slaney [33] emphasise the role of sequences for music similarity applications, whereas Aucouturier *et al.* [29] discuss the relative limitations of the bag-of-features approach in a comparison of musical and non-musical audio modelling. Sequential approaches have been utilized in music structure analysis, for identifying repeated and contrasting sequences and their boundaries within a single piece of music [34], in addition to cover song identification.

B. Cover Song Identification

Owing to the importance of tonal content in determining whether a song is a cover of another, recent cover song identification approaches typically extract representations of the tonal content using chroma features [35], [36]. Chroma features quantify energy distributions across octave-folded bands, using pitch classes in the chromatic scale to map frequency bands to chroma bins.

A variety of cover song identification approaches are based on aligning feature sequences. A widespread approach involves using dynamic programming to determine an optimal set of feature vector insertions, deletions and substitutions, obtained from a similarity matrix. Following Foote’s [37] method of applying dynamic time warping (DTW) to a similarity matrix constructed from spectral energy features, Gómez and Herrera [38] propose a DTW approach using chroma features. Serrà *et al.* [7] propose to compute binarized similarity matrices, substituting DTW with an alternative local alignment approach. The cross-recurrence approaches proposed by Serrà *et al.* [39] extend the notion of similarity matrices considered in the preceding investigations, in that time-lagged chroma vectors are combined to form higher-dimensional temporal features. In an alternative approach, Serrà *et al.* [40] utilise the previously described method of representing chroma features in combination with non-linear time series prediction techniques, using the cross-prediction error as a measure of similarity.

Using a signal processing approach, Ellis and Poliner [41] determine component-wise cross-correlation maxima as a measure of similarity between chroma features. Jensen [42] computes the Euclidean distance between two-dimensional autocorrelations of chroma sequences. More recently, Bertin-Mahieux [43] proposes a key-invariant approach based on applying the two-dimensional Fourier transform to chroma sequences.

An alternative approach involves computing similarities between discrete-valued representations of musical content. Tsai *et al.* [44] apply DTW to discrete-valued sequences, using spectral peak-picking for predominant melody extraction. Bello [45] and Lee [46] perform chord estimation with hidden Markov models, using mappings of model states to chords. The resulting sequences are then aligned using DTW. Martin *et al.* [47] heuristically select chroma bin maxima to determine triads, before locally aligning sequences. We may consider DTW-based approaches, the string-based heuristic evaluated in [47] and the cross-correlation approach evaluated in [41] as alignment techniques, in the sense that they may be used to maximise pairwise correlation between sequences.

With particular regard to this work, a number of approaches are based on applying the NCD to discrete-valued sequences. Using symbolic musical representations directly, Cilibrasi *et al.* [20] apply hierarchical clustering to pairwise distances between pieces of music, performing an analysis of clusters with respect to musical genres, musical works and artists. Li and Sleep apply the NCD to genre classification of symbolic musical representations [19] and musical audio [21].

For audio-based cover song identification, Ahonen [23] obtains discrete-valued representations of frame-based chroma features by applying a hidden Markov model (HMM) to perform chord transcription. Predicted chord sequences are then converted to a differential representation, before computing pairwise distances between tracks using the NCD based on different compression algorithms. Ahonen [48] further proposes to compute multiple discrete-valued representations using additional HMMs and by computing chroma differentials, before averaging separately obtained pairwise distances using the NCD based on prediction by partial matching (PPM) [49]. In addition, Ahonen [50] investigates chroma-derived representations which are compressed using Burrows-Wheeler (BW) compression [51]. Bello [24] applies the NCD to recurrence plots computed on individual tracks, as a measure of structural similarity between pieces of music. Finally, Tabus *et al.* [52] proposes a similar approach to Ahonen based on quantizing chroma-derived representations, observing that an alternative compression-based similarity measure outperforms the NCD. Additionally, Silva *et al.* [53] propose a measure of structural similarity based on video compression, observing superior performance using an alternative compression-based measure. Our work extends the above investigations, in that we examine and propose the use of alternative information-theoretic similarity measures to the NCD. Furthermore, we perform an extensive comparison of methods for estimating the NCD and related similarity measures, while proposing approaches which do not require quantizing audio features.

A number of recent investigations are concerned with cover song identification using large-scale music collections containing millions of tracks. For such collections, it is typically infeasible to perform computationally expensive pairwise comparisons between a query and every track in the collection. Casey *et al.* [54] compute Euclidean distances between windowed chroma sequences. Pairwise similarity is then quantified as the number of distances falling below a threshold. Such an approach may be combined with locality-sensitive hashing [55] for retrieval with sub-linear time complexity, with respect to a single query. Using a similar approach, Bertin-Mahieux and Ellis [56] propose to identify salient 'landmark' chroma vectors in individual tracks by applying a thresholding scheme. Identified landmark vectors are then encoded as an integer, thus the collection may be represented as a lookup table. Given a query, the same authors envisage that obtained results are re-ranked using a computationally expensive approach, as proposed by Khadkevich and Omologo [57]. In this work, we apply such a *filter-and-refine* approach [58], using information-theoretic similarity measures in the refinement stage.

C. Information-Theoretic Methods

Information-theoretic similarity measures between time series have been proposed in a variety of domains. The idea of jointly compressing two discrete-valued sequences is due to Loewenstern *et al.* [59] in the context of nucleotide sequence clustering. By parsing sequences using the Lempel-Ziv (LZ) algorithm [60], Ziv and Merhav [61] propose a method for comparing sequences by compressing one sequence using a model estimated on the other sequence. An alternative approach is considered by Benedetto *et al.* [62] for building language trees, where sequences are jointly compressed. Cilibrasi *et al.* [63] motivate their approach of jointly compressing sequences as an approximation of the normalized information distance [15].

III. Approach

We denote with $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M)$ two multivariate time series, each representing a sequence of feature vectors extracted from a piece of musical audio. If we assume that both $\mathbf{X}, \mathbf{Y}$ consist of independent and identically distributed realizations generated respectively by stochastic processes X, Y, one possible means of quantifying dissimilarity between time series involves the Kullback-Leibler (KL) divergence, defined as

$$D_{\mathrm{KL}}(p_X \| p_Y) = \int p_X(\mathbf{u}) \log\left(\frac{p_X(\mathbf{u})}{p_Y(\mathbf{u})}\right) d\mathbf{u} \quad (1)$$

where $p_X(\mathbf{u})$, $p_Y(\mathbf{u})$ denote the probability density of observation $\mathbf{u}$ emitted by X, Y, respectively. Viewed in terms of Shannon information and taking the logarithm to base 2, recall that the KL divergence quantifies the expected number of additional bits required to represent observations emitted by information source X, given an optimal code for observations emitted by information source Y. The KL divergence has been widely used in conjunction with a 'bag-of-features' approach for low-specificity music content analysis tasks [1].

To account for temporal structure in musical audio, we may use the NCD as a measure of musical dissimilarity between sequences of quantised feature vectors [21], [23], [52]. Given two strings $x = (x_1, x_2, \ldots, x_N)$, $y = (y_1, y_2, \ldots, y_M)$, the NCD is defined as

$$\mathrm{NCD}(x, y) = \frac{\max\{C(xy) - C(x), C(yx) - C(y)\}}{\max\{C(x), C(y)\}} \quad (2)$$

where $C(\cdot)$ denotes the number of bits required to encode a given string, using a compressor such as the LZ algorithm [60]. Similarly, $C(xy)$ denotes the number of bits required to encode the sequential concatenation of strings x, y. The NCD is an approximation of the normalized information distance (NID) [15], defined as

$$\mathrm{NID}(x, y) = \frac{K(x, y) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}} \quad (3)$$

where the uncomputable function $K(\cdot)$ denotes *algorithmic information content* (AIC), also known as Kolmogorov complexity. The AIC of a given string is the length in bits of the shortest program which outputs the string and then terminates [14]. Similarly, $K(x, y)$ denotes the length of the shortest program which outputs x, y, in addition to a means of distinguishing between both output strings [14]. Thus, AIC quantifies the number of bits required to represent specified input strings, under maximally attainable compression. Furthermore, the NID characterizes dissimilarity using the transformation under which input strings most closely resemble each other [15].

We are interested in examining the performance of the NCD as an approximation of the NID, where the choice of compressor determines the feature space used to compute similarities [64] in the NCD. Furthermore, note that the choice of sequential concatenation in $C(xy)$ to approximate $K(x, y)$ represents an additional heuristic [15]. In the following sections, we describe our contribution: We first consider in Section III-A the NID from the perspective of Shannon information, using which we propose a modification to the NCD in Section III-B. We then propose alternative prediction-based measures of similarity in Section III-C. We detail our approach of applying such measures to continuous-valued sequences in Section III-D.

A. Quantifying Time Series Dissimilarity Using Shannon Information

We approach the problem of quantifying dissimilarity from the perspective of Shannon information. We assume finite-order, stationary Markov sources X, Y. We denote with $X_{1:N}$ the sequence of discrete random variables $(X_1, \ldots, X_N)$ emitted by source X at times $1, \ldots, N$. We denote with $H_\mu(X)$, $H_\mu(X, Y)$, $H_\mu(X|Y)$ the entropy rate, joint entropy rate and conditional entropy rate, respectively defined as

$$H_\mu(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \quad (4)$$

$$H_\mu(X, Y) = \lim_{n\to\infty} \frac{1}{n} H((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)) \quad (5)$$

$$H_\mu(X|Y) = H_\mu(X, Y) - H_\mu(Y). \quad (6)$$

The entropy rate $H_\mu(X)$ defined in (4) quantifies the average amount of uncertainty about X_n, while accounting for dependency between X_n for all n. Analogously, the joint entropy rate $H_\mu(X, Y)$ defined in (5) quantifies the average amount of uncertainty about the pair (X_n, Y_n) emitted by sources X, Y, while in addition accounting for correlation between the sources. For the conditional entropy rate $H_\mu(X|Y)$ we have

$$H_\mu(X|Y) = \lim_{n\to\infty} \frac{1}{n} H(X_{1:n}, Y_{1:n}) - H(Y_{1:n}) \tag{7}$$

$$= \lim_{n\to\infty} \frac{1}{n} H(X_{1:n}|Y_{1:n}). \tag{8}$$

From (8) we may interpret $H_\mu(X|Y)$ as quantifying the average amount of uncertainty about a given emission X_n, while taking into account dependency between observations emitted by X and given knowledge of observations emitted by Y.

For N observations emitted from source X, up to an additive constant the expectation $\mathbf{E}[K(X_{1:N})]$ may be approximated using the entropy [65],

$$\mathbf{E}[K(X_{1:N})] \approx H(X_{1:N}). \tag{9}$$

Using (4), (5), we assume further approximations

$$\mathbf{E}[K(X_{1:N})] \approx NH_\mu(X) \tag{10}$$

$$\mathbf{E}[K(X_{1:N}, Y_{1:N})] \approx NH_\mu(X, Y) \tag{11}$$

where $\mathbf{E}[K(X_{1:N}, Y_{1:N})]$ denotes the expected value of $K(\cdot, \cdot)$ for N observations emitted from sources X, Y. In terms of Shannon information, following [66] we use (6) and estimate the NID as

$$\mathrm{NID}(X, Y) \approx \frac{\max\{H_\mu(X|Y), H_\mu(Y|X)\}}{\max\{H_\mu(X), H_\mu(Y)\}}. \tag{12}$$

B. Normalized Compression Distance with Alignment

As given in (12), the NID utilizes the joint entropy rate $H_\mu(X, Y)$, which accounts for correlation between sources. In contrast, the approach of compressing sequentially concatenated strings to estimate $K(x, y)$ may be inadequate for compressors based on Markov sources, since correlation is not accounted for [66]. To address this possible limitation, we propose the normalized compression distance with alignment (NCDA), defined as

$$\mathrm{NCDA}(x, y) = \frac{C(\langle x, y\rangle) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \tag{13}$$

where $\langle a, b\rangle$ performs alignment as a means of maximizing correlation between integer-valued strings a, b. We generate equal-length strings by padding the shorter of the two strings with the most common value of the longer string. Then, we determine the lag which maximizes cross-correlation between strings, before circularly shifting b using the obtained lag value. Finally, we interleave strings. We motivate our choice of cross-correlation by considering that cross-correlation may be computed efficiently, as a series of inner products. Hence, our choice of cross-correlation is pragmatic; an alternative approach might involve minimizing NCDA with respect to all lags, or aligning strings using an alternative algorithm.

C. Predictive Modelling

As previously described, the NCD and NCDA rely on determining the number of bits required to encode strings, using a specified compression algorithm. As an alternative approach, we consider the relation between predictability and compressibility [67], [68] and perform sequence prediction. We illustrate our approach for the case of discrete-valued observations. First, recall that the entropy rate $H_\mu(X)$ is given as

$$H_\mu(X) = \lim_{n\to\infty} -\frac{1}{n} \sum_{x_{1:n}\in\mathcal{A}^n} P_X(x_{1:n}) \log P_X(x_{1:n}) \tag{14}$$

where $P_X(x_{1:n})$ denotes the probability of observing $X_{1:n} = x_{1:n}$, with $x_{1:n} \in \mathcal{A}^n$ according to the alphabet $\mathcal{A}$. We may interpret the quantity $-\log P_X(x_{1:n})$ as the number of bits required to represent $\mathbf{u}_{1:n}$, assuming an optimal code. $H_\mu(X)$ thus quantifies the expected number of bits required to represent a single observation emitted by X, while accounting for dependency between observations. Assume that we have an empirical estimate $\hat{P}_X$ of the distribution P_X, based on finite observations $x_{1:N}$. Following [69], we estimate $H_\mu(X)$ using *average log-loss* $\ell(\hat{P}_X, x_{1:N})$, defined as

$$\ell(\hat{P}_X, x_{1:N}) = -\frac{1}{N} \log \hat{P}_X(x_{1:N}) \tag{15}$$

$$= -\frac{1}{N}\left(\log \hat{P}_X(x_1) + \sum_{i=2}^{N} \log \hat{P}_X(x_i|x_{1:i-1})\right) \tag{16}$$

where $\hat{P}_X(x_i|x_{1:i-1})$ denotes the estimated probability of observing x_i, given preceding context $x_{1:i-1}$. Using (16), we thus compute average log-loss by evaluating the likelihood of observations $x_{1:i-1}$ under the estimated distribution $\hat{P}_X$, which we may conceive of as performing a series of predictions based on increasingly long contexts $x_{1:i-1}$. Since $\hat{P}_X$ is an estimate of P_X, the described process is termed *self-prediction* [40].

We denote with $P_Y(x_{1:n})$ the probability of observing $x_{1:n}$ from source Y. A measure of disparity between sources X, Y is the cross entropy rate $H_\mu^\times(X, Y)$,

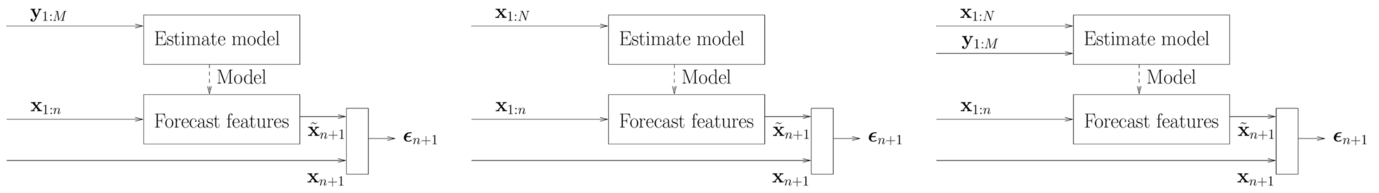
$$H_\mu^\times(X, Y) = \lim_{n\to\infty} -\frac{1}{n} \sum_{x_{1:n}\in\mathcal{A}^n} P_X(x_{1:n}) \log P_Y(x_{1:n}) \tag{17}$$

quantifying the expected number of bits required to represent observations emitted by source X, given an optimal code for source Y. We estimate $H_\mu^\times(X, Y)$ by computing the average log-loss $\ell(\hat{P}_Y, x_{1:N})$ based on iterated prediction, where $\hat{P}_Y$ denotes an estimate of P_Y based on observations $y_{1:M}$. Since $\hat{P}_Y$, $\hat{P}_X$ represent disparate sources, the described process is termed *cross-prediction* [40]. Analogous to the NCD, as a symmetric distance between sources X, Y based on cross entropy, we compute the quantity

$$D^\times(X, Y) = \frac{H_\mu^\times(X, Y) + H_\mu^\times(Y, X)}{H_\mu(X) + H_\mu(Y)} \tag{18}$$

where in (18) the denominator serves as a normalization factor, analogous to the denominator in (2) and where we use self-prediction to estimate $H_\mu(X), H_\mu(Y)$.

To obtain a prediction-based estimate of the NID in (12), we may estimate $H_\mu(X)$, $H_\mu(Y)$ again using self-prediction. Furthermore, we estimate the conditional entropy rate $H_\mu(X|Y)$



(a) Cross-prediction (b) Self-prediction (c) Conditional self-prediction

Fig. 1. Evaluated prediction strategies. Sequences $\mathbf{x}_{1:N}, \mathbf{y}_{1:M}$ serve as model inputs, observation context $\mathbf{x}_{1:n}$ forms basis of prediction $\tilde{\mathbf{x}}_{n+1}$. Quantity ϵ_{n+1} denotes prediction error.

using the distribution $\hat{P}_{X|Y}$, referring to the estimated distribution of observations emitted by X, given knowledge of observations $y_{1:M}$ emitted by Y. Analogous to self-prediction and cross-prediction, we define the quantity $\ell(\hat{P}_{X|Y}, x_{1:N}, y_{1:M})$,

$$\ell(\hat{P}_{X|Y}, x_{1:N}, y_{1:M}) = -\frac{1}{N}\left(\log \hat{P}_{X|Y}(x_1|y_{1:M}) + \sum_{i=2}^{N} \log \hat{P}_{X|Y}(x_i|x_{1:i-1}, y_{1:M})\right). \tag{19}$$

We refer to the process used to compute (19) as *conditional self-prediction*.

D. Continuous-Valued Approach

The quantities described in Section III-C may be computed using quantised feature vectors [21], [23], [31], [52]. As an alternative, we propose an approach requiring no prior quantisation. As used in [40], in our approach we utilise non-linear time series prediction. In contrast to [40], we are concerned with evaluating distance measures which we compute as statistics of prediction errors. Therefore, we use a comparatively straightforward nearest-neighbors approach. Given the sequence of feature vectors $\mathbf{C}$, consider first the process of *time-delay embedding* [70], which yields the vector sequence $\mathbf{S}^{\mathbf{C}}$, whose elements $\mathbf{s}_r^{\mathbf{C}}$ are defined as

$$\mathbf{s}_r^{\mathbf{C}} = \mathrm{vec}(\mathbf{c}_r, \mathbf{c}_{(r-1)\tau}, \ldots, \mathbf{c}_{(r-d+1)\tau}). \tag{20}$$

According to (20), each element $\mathbf{s}_r^{\mathbf{C}}$ aggregates feature vector $\mathbf{c}_r$ along with its preceding temporal context $(\mathbf{c}_{(r-1)\tau}, \ldots, \mathbf{c}_{(r-d+1)\tau})$. The amount of temporal context is controlled by parameters d, τ, respectively referred to as *embedding dimension* and *time delay*. Operator vec denotes vectorization.

Our method of predicting features is based on determining nearest neighbors in time-delay embedded space. We first illustrate our method for the case of cross-prediction, depicted schematically in Fig. 1(a). Given sequence $\mathbf{y}_{1:M}$, we denote with $\tilde{\mathbf{x}}_{t+h}$ the estimated successor of sequence $\mathbf{x}_{1:t+h-1}$,

$$\tilde{\mathbf{x}}_{t+h} = \mathbf{y}_{q(t)+h} \tag{21}$$

where h denotes the *predictive horizon* (how far into the future we predict), and where we define $q(t)$ as

$$q(t) = \operatorname*{arg\,max}_{k \in [d..M-h]} \mathrm{corr}(\mathbf{s}_k^{\mathbf{Y}}, \mathbf{s}_t^{\mathbf{X}}) \tag{22}$$

with $\mathrm{corr}(\mathbf{s}_k^{\mathbf{Y}}, \mathbf{s}_t^{\mathbf{X}})$ denoting the sample Pearson correlation coefficient between vectors $\mathbf{s}_k^{\mathbf{Y}}, \mathbf{s}_t^{\mathbf{X}}$. We motivate use of correlation coefficients as an alternative to the Euclidean distance, following [71].

Depicted schematically in Fig. 1(b), to perform self-prediction we set $\mathbf{Y} = \mathbf{X}$. Since features may be slowly-varying, when forming prediction $\tilde{\mathbf{x}}_{t+h}$ we disregard observations in the immediate past of time step t. Thus we define

$$\tilde{\mathbf{x}}_{t+h} = \mathbf{x}_{q'(t)+h} \tag{23}$$

with $q'(t)$ defined as

$$q'(t) = \operatorname*{arg\,max}_{k \in [d..N-h], |k-t|>R} \mathrm{corr}(\mathbf{s}_k^{\mathbf{X}}, \mathbf{s}_t^{\mathbf{X}}) \tag{24}$$

and where R denotes the radius below which observations are disregarded.

Finally, to perform conditional self-prediction, we use both time-delay embedded spaces $\mathbf{s}^{\mathbf{Y}}$, $\mathbf{s}^{\mathbf{X}}$. Given predictions $\mathbf{y}_{q(t)+h}$, $\mathbf{x}_{q'(t)+h}$, respectively obtained using cross-prediction and self-prediction, we compute the linear combination

$$\tilde{\mathbf{x}}_{t+h} = \mathbf{y}_{q(t)+h}\alpha + \mathbf{x}_{q'(t)+h}(1-\alpha). \tag{25}$$

Similar to the approach given in [72], in (25) for weighting coefficient α we use

$$\alpha = \frac{\mathrm{MSE}_{\mathrm{self}}}{\mathrm{MSE}_{\mathrm{self}} + \mathrm{MSE}_{\mathrm{cross}}} \tag{26}$$

where $\mathrm{MSE}_{\mathrm{cross}}$, $\mathrm{MSE}_{\mathrm{self}}$ respectively denote cross-prediction and self-prediction mean squared errors. Fig. 1(c) depicts conditional self-prediction schematically.

Given the sequence of predictions $\tilde{\mathbf{x}}_{1:N}$, we denote with $\boldsymbol{\epsilon}_n$ the rescaled prediction error, whose ith component $\epsilon_{i,n}$ is given by

$$\epsilon_{i,n} = \frac{\tilde{x}_{i,n} - x_{i,n}}{s_i} \tag{27}$$

where s_i denotes the sample variance of the ith component $(\mathbf{x}_{1:N})_i$ in $\mathbf{x}_{1:N}$. We contrast our approach with the component-wise normalized mean squared error (NMSE) based on cross-prediction used in [40], which may be applied as an alternative measure of dissimilarity between time series. Our approach is based on assuming that the prediction error may be represented using a normally distributed random variable Z with samples $\boldsymbol{\epsilon}_{1:N}$. Using the samples, we estimate the prediction error entropy $H(Z)$ parametrically. In the case of self-prediction, we assume the approximation $H(Z) \approx H_\mu(X)$; analogously in the case of cross-prediction and conditional self-prediction, we assume respective approximations $H(Z) \approx H^{\times}_{\mu}(X,Y), H(Z) \approx H_{\mu}(X|Y)$. Assuming normality, we estimate $H(Z)$ using the equation

$$H(Z) = \frac{1}{2}\log{(2\pi e)^k|\mathbf{\Sigma}|} \tag{28}$$

where $\mathbf{\Sigma}$ denotes the sample covariance of Z. In our continuous-valued approach, using the prediction methods depicted in Fig. 1, we thus estimate information-based measures as statistics of the prediction error sequence. We then substitute the obtained quantities in (12) and (18) to obtain continuous-valued, prediction-based analogues of the NID and distance $D^{\times}$. The continuous-valued, prediction-based approach contrasts with our discrete-valued, prediction-based methods previously described in Section III-C and our discrete-valued, compression-based method described in Section III-B.

IV. Experimental Method

We first evaluate our proposed methods using a set of 300 audio recordings of Jazz standards[1]. We assume that two tracks are a cover pair if they possess identical title strings. Thus, we assume a symmetric relation when determining cover identities. The equivalence class of tracks deemed to be covers of one another is a *cover set*. The Jazz data set comprises 97 cover sets, with average cover set size 3.06 tracks.

Furthermore, we perform a large-scale evaluation based on the MSD [25]. This dataset includes meta-data and pre-computed audio features for a collection of 10^6 Western popular music recordings. We use a pre-defined evaluation set of 5236 query tracks partitioned into 1726 cover sets[2], with average cover set size 3.03 tracks. Following [43], for each query track, we seek to identify the remaining cover set members contained in the entire 10^6 track collection.

A. Feature Extraction

For the Jazz dataset, as a representation of musical harmonic content, we extract 12-component beat-synchronous chroma features from audio using the method and implementation described in [41]. Assuming an equal-tempered scale, the method accounts for deviations in standard pitch from 440 Hz, by shifting the mapping of FFT bins to pitches in the range of ± 0.5 semitones. Following chroma extraction, beat-synchronisation is achieved using the method described in [73]. First, onset detection is performed by differencing a log-magnitude Mel-frequency spectrogram across time and applying half-wave rectification, before summing across frequency bands. After high-pass filtering the onset signal, a tempo estimate is formed by applying a window function to the autocorrelated onset signal and determining autocorrelation maxima. Varying the centre of the window function allows tempo estimation to incorporate a bias towards a preferred beat rate (PBR). The tempo estimate and onset signal are then used to obtain an optimal set of beat onsets, by using dynamic programming. Chroma features are averaged over beat intervals, before applying square-root compression and normalizing chroma features with respect to the Euclidean norm. Based on our previous work [13], we evaluate using a PBR of 240 beats per minute (bpm).

The MSD includes 12-component chroma features alongside predicted note and beat onsets [74], which we use in our evaluations. In contrast to the beat-synchronous features obtained for the Jazz dataset, MSD chroma features are initially aligned to predicted onsets. Motivated by our choice of PBR for the Jazz dataset, we resample predicted beat onsets to a rate of 240 bpm. We then average chroma features over resampled beat intervals. Finally, we normalize features as described for the Jazz dataset.

B. Key Invariance

To account for musical key variation within cover sets, we transpose chroma sequences using the optimal transposition index (OTI) method [7]. Given two chroma vector sequences $\mathbf{X}$, $\mathbf{Y}$, we form summary vectors $\mathbf{h_X}$, $\mathbf{h_Y}$ by averaging over entire sequences. The OTI corresponds to the number of circular shift operations applied to $\mathbf{h_Y}$ which maximizes the inner product between $\mathbf{h_X}$ and $\mathbf{h_Y}$,

$$\mathrm{OTI}(\mathbf{h_X},\mathbf{h_Y}) = \arg\max_{i} \mathbf{h_X}\cdot \mathrm{circshift}(\mathbf{h_Y},i) \tag{29}$$

where $\mathrm{circshift}(\mathbf{h_Y},i)$ denotes applying i circular shift operations to $\mathbf{h_Y}$. We subsequently shift chroma vectors $\mathbf{Y}$ by $\mathrm{OTI}(\mathbf{h_X},\mathbf{h_Y})$ positions, prior to pairwise comparison.

C. Quantisation

For discrete-valued similarity measures, we quantize chroma features using the K-means algorithm. We cluster chroma features aggregated across all tracks, where we consider codebook sizes in the range [2..48]. To increase stability, we execute the K-means algorithm 20 times. We then select the clustering which minimizes the mean squared error between data points and assigned clusters. The described quantisation method performs similarly to an alternative based on pairwise sequence quantisation; for a detailed discussion we refer to our previous work [13].

D. Distance Measures

We summarize the distance measures evaluated in this work in Table I, where for each distance measure, we list our estimation methods.

We utilise the following algorithms to compute distance measures by compressing strings: Prediction by partial matching (PPM) [49], Burrows-Wheeler (BW) compression [51] and Lempel-Ziv (LZ) compression [60], implemented respectively as PPMD[3], BZIP2[4] and ZLIB[5]. In all cases, we set parameters to favour compression rates over computation time. To obtain strings, following quantisation we map integer codewords to alphanumeric characters.

We use the described compression algorithms to determine the length in bits of compressed strings and compute NCD,

[1] http://www.eecs.qmul.ac.uk/~peterf/jazzdataset.html

[2] http://labrosa.ee.columbia.edu/millionsong/secondhand

[3] http://compression.ru/ds/

[4] http://bzip2.org

[5] http://zlib.org

TABLE I
SUMMARY OF EVALUATED DISTANCE MEASURES

Distance measure	Definition	Estimation method	
NCD	Eqn. 2	String compression (LZ, BW, PPM)	Discrete prediction (LZ, PPM)
NCDA	Eqn. 13	String compression (LZ, BW, PPM)	Discrete prediction (LZ, PPM)
$D^{\times}$	Eqn. 18	Discrete prediction	Continuous prediction
D_{JS}	Eqn. 31	Normalised symbol histograms	
NID	Eqn. 12	Continuous prediction	

NCDA distances. In a complementary discrete-valued approach, we use string prediction instead of compression. Using average log-loss, we compute NCDA using the formula

$$\frac{\ell(\hat{P}_{\langle X,Y\rangle},\langle x,y\rangle)-\min\{\ell(\hat{P}_X,x),\ell(\hat{P}_Y,y)\}}{\max\{\ell(\hat{P}_X,x),\ell(\hat{P}_Y,y)\}} \tag{30}$$

where $\ell(\hat{P}_{\langle X,Y\rangle},\langle x,y\rangle)$ is the average log-loss obtained from performing self-prediction on the aligned sequence $\langle x,y\rangle$. We compute a prediction-based variant of NCD analogously by predicting sequentially concatenated strings without performing any alignment. In addition, we use cross-prediction to estimate distance measure $D^{\times}$, as defined in (18). We perform string prediction using Begleiter's [69] implementations of PPMC and LZ78 algorithms.

Note that the KL divergence given in (1) is non-symmetric. In our evaluations, we observed that computing a symmetric distance improved performance; based on KL divergence, we compute the Jensen-Shannon divergence (JSD) $D_{\mathrm{JS}}(p_X \| p_Y)$, defined as

$$D_{\mathrm{JS}}(p_X \| p_Y) = D_{\mathrm{KL}}(p_X \| p_A) + D_{\mathrm{KL}}(p_Y \| p_A) \tag{31}$$

where p_A denotes the mean of p_X, p_Y,

$$p_A = \frac{1}{2}\left(p_X + p_Y\right). \tag{32}$$

As a baseline method, we compute the JSD between symbol histograms normalized to sum to one.

We evaluate continuous-valued prediction using time-delay embedding parameters $h \in \{1,4\}$, $d \in \{1,2,4\}$, $\tau \in \{1,2,4,6\}$, setting the exclusion radius in (24) to $R = 8$ based on preliminary analysis using separate training data. We compute distance measure $D^{\times}$ using cross-prediction to estimate the numerator in (18). In a complementary approach, we estimate the NID using conditional self-prediction to estimate the numerator in (12). For $D^{\times}$ and NID, we use self-prediction to estimate the denominator in (18), (12), respectively.

Finally, to compensate for cover song candidates consistently deemed similar to query tracks, we normalize pairwise distances using the method described in [75]. We apply distance normalization as a post-processing step, before computing performance statistics.

E. Large-scale Cover Song Identification

For music content analysis involving large datasets, algorithm scalability is an important issue. The approaches in this work by themselves require a linear scan through the dataset for a given query, which may be infeasible for large datasets. We use a scalable approach for our evaluations involving the MSD. Following [57] and similar to the method proposed in [76], we incorporate our methods into a two-stage retrieval process. By using a metric distance to determine similarity in the first retrieval stage, we allow for the potential use of indexing or hashing schemes, as proposed in [54], [58]. We then apply non-metric pairwise comparisons in the second retrieval stage.

In the first stage, we quantize as described in Section IV-C and represent each track with a normalized codeword histogram. Given a query track, we then rank each of the 10^6 candidate tracks using the L1 distance. To account for key variation, for each candidate track we minimise L1 distance across chroma rotations. We then determine the top $L = 1000$ candidate tracks, which we re-rank in the second stage using our proposed methods. After both retrieval stages, we normalize pairwise distances as described in Section IV-D. We report performance based on the final ranking of all 10^6 candidate tracks, across query tracks.

F. Performance Statistics

As used in [24], we quantify cover song identification accuracy using mean average precision (MAP), based on ranking tracks according to distance with respect to queries. The MAP is obtained by averaging query-wise scores, where we may interpret each score as the average of precision values at the ranks of relevant tracks, where relevant tracks in our case are covers of the query track. Following [24], we use the Friedman test [77] to compare accuracies among distance measures. The Friedman test is based on ranking across queries each distance measure according to average precision. We combine the Friedman test with Tukey's range test [78] to adjust for Type I errors when performing multiple comparisons.

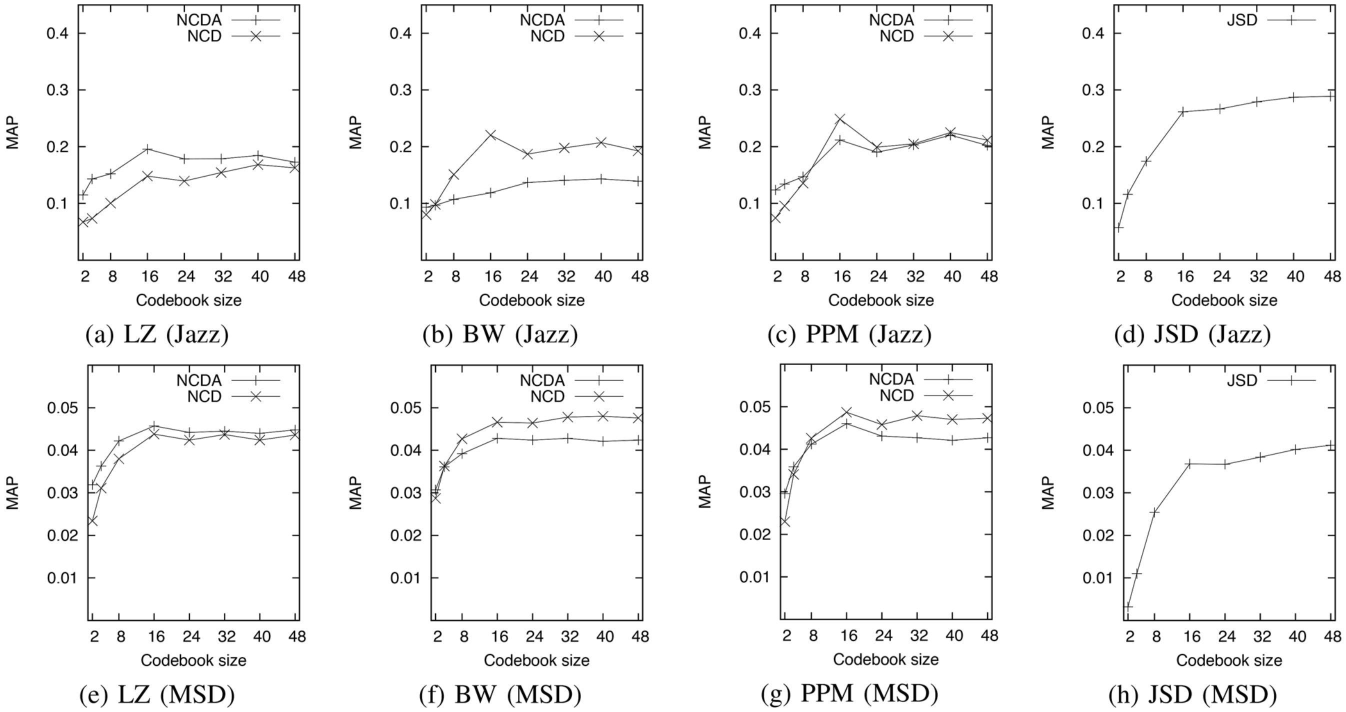
As a subsidiary performance measure, for each query we compute the precision at rank r, with $r \in \{5, 10, 20\}$. We subsequently average across queries to obtain mean precision at rank r.

G. Combining Distance Measures

To determine if combining distance measures improves cover song identification accuracy, we obtain pairwise distances as described in Section IV-D. We denote with $d^k_{i,j}$ the pairwise distance between the ith query track and the jth result candidate, obtained using the kth distance measure in our evaluation. We transform $d^k_{i,j}$ by computing the inverse rank ${d'}^{k}_{i,j}$,

$${d'}^{k}_{i,j} = 1 - \mathrm{rank}(d^k_{i,j})^{-1} \tag{33}$$

where $\mathrm{rank}(d^k_{i,j})$ denotes the rank of $d^k_{i,j}$ among all distances obtained with respect to query track i, given the kth distance measure. We apply this transformation to protect against outliers, while ensuring that distance decreases rapidly for track pairs deemed highly similar, for decreasing distance. Note that since our distance transformation preserves monotonicity and



(a) LZ (Jazz)

(b) BW (Jazz)

(c) PPM (Jazz)

(d) JSD (Jazz)

(e) LZ (MSD)

(f) BW (MSD)

(g) PPM (MSD)

(h) JSD (MSD)

Fig. 2. Effect of codebook size and distance measure on mean average precision (MAP). Results displayed for Lempel-Ziv (LZ), Burrows-Wheeler (BW) and prediction by partial matching (PPM) algorithms in subfigures (a)–(c), (e)–(g), for Jazz and MSD datasets respectively. Subfigures (d), (h) display results for Jensen-Shannon divergence baseline (JSD), for Jazz and MSD datasets respectively.

MAP itself is based on ranked distances, performance of unmixed distance measures is uninfluenced by this transformation. Finally, we combine distances ${d'}^{k}_{i,j}$, ${d'}^{m}_{i,j}$ by computing a weighted average of distances pooled using max and min operators,

$$\max\{{d'}^{k}_{i,j}, {d'}^{m}_{i,j}\}\beta + \min\{{d'}^{k}_{i,j}, {d'}^{m}_{i,j}\}(1-\beta) \qquad (34)$$

where we vary β in the range [0, 1]. We motivate our approach on the basis that we may interpret inverse ranks as estimated probabilities of cover identities, furthermore the operators max and min have been proposed as a means of combining probability estimates for classification [79]. In forming a linear combination, we evaluate the utility of max pooling versus min pooling. An alternative approach based on straightforward averaging did not yield any performance gain.

H. Baseline Approaches

In addition to the JSD and cross-prediction NMSE baselines, we include an evaluation of the method and implementation described in [41] based on cross-correlation. As a random baseline, we sample pairwise distances from a normal distribution.

V. Results

A. Discrete-Valued Approaches Based on Compression

In Fig. 2(a)–(c), we examine the performance of discrete-valued NCD and NCDA distance measures, combined with LZ, BW and PPM algorithms and based on the Jazz dataset. For the LZ algorithm, NCDA yields a relative performance gain of 38.6%, averaged across codebook sizes. In contrast, for PPM, with the exception of small codebook sizes in the range [2..8], NCDA yields no consistent improvement over NCD, however averaged across codebook sizes we obtain a mean relative performance gain of 11.0%. Finally, the effect of using NCDA is reversed for BW compression, where performance decreases by an average of 21.8%.

Examining results for the MSD in Fig. 2(e)–(g), we observe similar qualitative results for LZ and BW algorithms. For the LZ algorithm, NCDA yields an average relative performance gain of 10.1%, whereas for BW compression we observe an average relative performance loss of 6.5%. In contrast to the Jazz dataset, for PPM we observe an average relative performance loss of 1.5%.

For both datasets, NCDA appears to be most advantageous combined with LZ compression, whereas BW yields the least advantageous result. Note that BW compression is block-based in contrast to LZ and PPM compressors, both of which are sequential. We attribute this observation to performance differences among compressors, since the assumptions made in Section III-B rely on assuming Markov sources. Noting differences in relative performance gains between datasets, following [57] we further conjecture that chroma feature representation influences the performance of the evaluated distance measures.

We examine the performance of JSD between normalized symbol histograms, as displayed in Fig. 2(d), (h). Surprisingly, for the Jazz dataset and for $K > 8$, JSD outperforms compression-based methods, with maximum MAP score 0.289 obtained for $K = 48$. This result is contrary to our expectation that NCD approaches should outperform the bag-of-features approach, by accounting for temporal structure in time series. In contrast, for the MSD and for optimal K, both NCD and NCDA outperform JSD across all evaluated compression algorithms. We attribute

TABLE II
MEAN AVERAGE PRECISION SCORES FOR DISTANCES BASED ON CONTINUOUS PREDICTION. IN EACH SUBFIGURE, PARAMETERS h, τ, d DENOTE PREDICTIVE HORIZON, TIME DELAY AND EMBEDDING DIMENSION, RESPECTIVELY. RESULTS DISPLAYED IN SUBFIGURES (A)–(C), (D)–(F) FOR JAZZ AND MSD DATASETS, RESPECTIVELY.

	d	$\tau = 1$	$\tau = 2$	$\tau = 4$	$\tau = 6$
h=1	1	0.282	0.282	0.282	0.282
	2	0.308	0.311	0.293	0.312
	4	0.327	0.332	0.318	0.318
h=4	1	0.243	0.243	0.243	0.243
	2	0.262	0.273	0.291	0.284
	4	0.307	0.313	**0.346**	0.321

(a) NID estimate; conditional self-prediction (Jazz)

	d	$\tau = 1$	$\tau = 2$	$\tau = 4$	$\tau = 6$
h=1	1	0.347	0.347	0.347	0.347
	2	0.412	0.403	0.390	0.403
	4	**0.454**	0.446	0.432	0.423
h=4	1	0.293	0.293	0.293	0.293
	2	0.352	0.364	0.377	0.365
	4	0.408	0.428	0.432	0.435

(b) $D^{\times}$ estimate; cross-prediction (Jazz)

	d	$\tau = 1$	$\tau = 2$	$\tau = 4$	$\tau = 6$
h=1	1	0.344	0.344	0.344	0.344
	2	0.402	0.396	0.385	0.389
	4	0.448	0.452	0.428	0.433
h=4	1	0.321	0.321	0.321	0.321
	2	0.362	0.375	0.390	0.379
	4	0.417	0.450	0.446	**0.459**

(c) NMSE; cross-prediction (Jazz)

	d	$\tau = 1$	$\tau = 2$	$\tau = 4$	$\tau = 6$
h=1	1	0.0191	0.0191	0.0191	0.0191
	2	0.0230	0.0222	0.0239	0.0250
	4	0.0238	0.0275	**0.0303**	0.0295
h=4	1	0.0200	0.0200	0.0200	0.0200
	2	0.0208	0.0239	0.0236	0.0260
	4	0.0228	0.0276	0.0303	0.0301

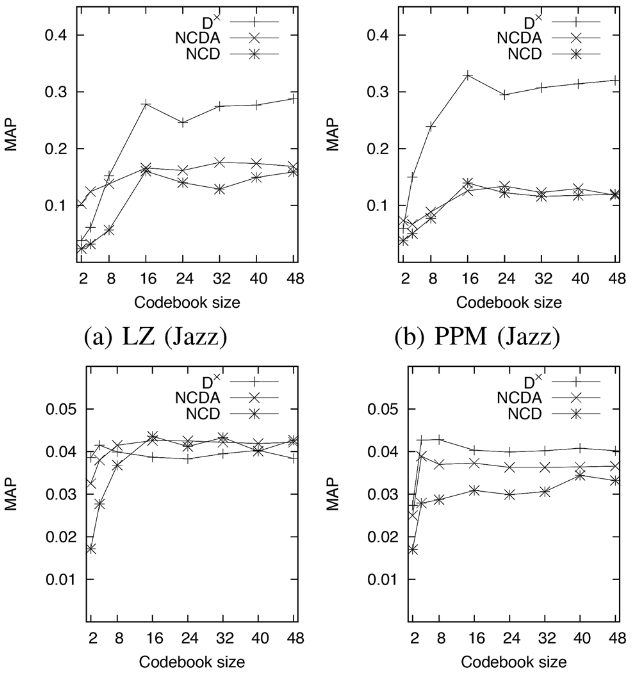
(d) NID estimate; conditional self-prediction (MSD)

	d	$\tau = 1$	$\tau = 2$	$\tau = 4$	$\tau = 6$
h=1	1	0.0451	0.0451	0.0451	0.0451
	2	0.0476	0.0477	0.0479	0.0475
	4	0.0489	0.0494	0.0494	0.0489
h=4	1	0.0465	0.0465	0.0465	0.0465
	2	0.0470	0.0480	0.0484	0.0487
	4	0.0478	0.0488	**0.0498**	0.0491

(e) $D^{\times}$ estimate; cross-prediction (MSD)

	d	$\tau = 1$	$\tau = 2$	$\tau = 4$	$\tau = 6$
h=1	1	0.0341	0.0341	0.0341	0.0341
	2	0.0404	0.0420	0.0431	0.0437
	4	0.0447	0.0474	0.0478	0.0465
h=4	1	0.0431	0.0431	0.0431	0.0431
	2	0.0450	0.0457	0.0467	0.0471
	4	0.0466	0.0494	**0.0499**	0.0494

(f) NMSE; cross-prediction (MSD)



(a) LZ (Jazz) (b) PPM (Jazz)

(c) LZ (MSD) (d) PPM (MSD)

Fig. 3. Effect of codebook size and distance measure on mean average precision (MAP). Results obtained using string prediction approach, displayed for Lempel-Ziv (LZ) (subfigures (a), (c)) and prediction by partial match (PPM) (subfigures (b), (d)), for Jazz and MSD datasets respectively.

this disparity to differences in dataset size, where for the Jazz dataset the problem size may be sufficiently small to amortize advantages of using NCD, NCDA compared to JSD.

B. Discrete-Valued Approaches Based on Prediction

In Fig. 3, we consider the performance of distance measures based on string prediction. For the Jazz dataset, comparing log-loss estimates of NCD and NCDA using the LZ algorithm, averaged across codebook sizes NCDA outperforms NCD; we obtain a mean relative performance gain of 105.1% (Fig. 3(a)). For the PPM algorithm, although NCD maximizes performance (MAP 0.140), we obtain a mean relative performance gain of 19.3% using NCDA over NCD (Fig. 3(b)). Importantly, for both LZ and PPM the cross-prediction distance $D^{\times}$ consistently outperforms NCD and NCDA; for $K = 16$ and combined with PPM compression, we obtain MAP 0.329. For the MSD and using LZ compression, in contrast to the Jazz dataset we observe a mean relative performance loss of 1.8% when comparing $D^{\times}$ with NCDA. For both LZ and PPM, NCDA compared to NCD yields mean relative performance gains of 17.6% and 24.0%, respectively.

C. Continuous-Valued Approaches

Table II displays the performance of continuous-valued prediction approaches. Note that for $d = 1$, parameter τ may be set to an arbitrary integer following (20). We consider results obtained for the Jazz dataset (Table II(a)–(c)). Using conditional self-prediction to estimate the NID, maximized across parameters h, d, τ we obtain MAP 0.346. In comparison, cross-prediction distance $D^{\times}$ yields MAP 0.454. As a baseline, we determine the cross-prediction NMSE, where maximizing across parameters we obtain MAP 0.459. Table II(a)–(c) displays performance against evaluated parameter combinations. Examining results for the MSD in Table II(d)–(f), we obtain qualitatively similar results with maximum MAP values 0.0303, 0.0498 and 0.0499 for NID, $D^{\times}$ and NMSE, respectively. For both datasets, we observe that increasing the value of d consistently improves performance. In contrast, we observe no such effect for parameters τ, h.

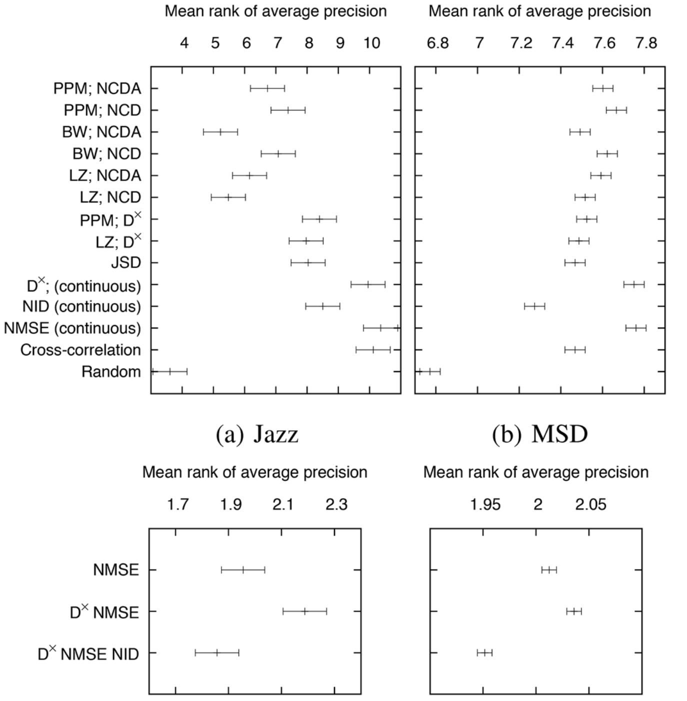
D. Summary of Results and Comparison to State of the Art

Fig. 4(a), (b) displays the result of significance testing as described in Section IV-F, where we assume 95% confidence intervals and where we maximise across evaluated parameter spaces. Table III displays a corresponding summary of MAP scores. As baselines we include Ellis and Poliner's cross-correlation approach [41], in addition to randomly sampled pairwise distances. For the MSD, when used without any further refinement method, our filtering stage based on normalized codeword histograms yields MAP 0.0056.

For both Jazz dataset and MSD, we observe that continuous-valued approaches based on cross-prediction consistently outperform discrete-valued approaches. Moreover, with the exception of NCD combined with PPM-based string compression

TABLE III
SUMMARY OF MEAN AVERAGE PRECISION SCORES. FIRST THREE ROWS DENOTE COMPRESSION BASED APPROACHES. INTERVALS ARE STANDARD ERRORS. 'RANDOM' DENOTES SAMPLING PAIRWISE DISTANCES FROM A NORMAL DISTRIBUTION

Dataset	Jazz		MSD	
Method	NCDA	NCD	NCDA	NCD
PPM	0.220 ± 0.021	0.249 ± 0.021	0.0460 ± 0.0024	0.0487 ± 0.0025
BW	0.143 ± 0.016	0.220 ± 0.019	0.0428 ± 0.0023	0.0480 ± 0.0024
LZ	0.196 ± 0.019	0.168 ± 0.017	0.0457 ± 0.0024	0.0438 ± 0.0023
PPM; $D^{\times}$	0.329 ± 0.022		0.0428 ± 0.0022	
LZ; $D^{\times}$	0.288 ± 0.021		0.0415 ± 0.0022	
JSD	0.289 ± 0.022		0.0412 ± 0.0023	
$D^{\times}$ (continuous)	0.454 ± 0.024		0.0498 ± 0.0025	
NID (continuous)	0.346 ± 0.023		0.0303 ± 0.0020	
NMSE (continuous)	0.459 ± 0.023		0.0499 ± 0.0025	
Ellis and Poliner [41]	0.465 ± 0.024		0.0404 ± 0.0023	
Random	0.026 ± 0.004		0.0006 ± 0.0001	
$D^{\times}$ & NMSE (cont.)	0.496		0.0516 ± 0.0025	
$D^{\times}$ & NID & NMSE (cont.)	0.432		0.0463 ± 0.0024	



(a) Jazz (b) MSD

(c) Jazz (d) MSD

Fig. 4. Mean ranks of average precision scores obtained using Friedman test. Error bars indicate 95% confidence intervals obtained using Tukey's range test [78]. Higher mean ranks indicate higher performance. Results displayed for Jazz and MSD datasets in subfigures (a) and (b), respectively, with results for combined distances displayed in subfigures (c) and (d).

and for the MSD, using continuous-valued cross-prediction significantly outperforms discrete-valued approaches. For approaches based on string compression, we note that using NCDA with BW compression significantly decreases performance with respect to NCD. Similarly, using NCDA decreases MAP scores for PPM. Although we do not observe a significant performance gain using NCDA over NCD for LZ compression, performance improves consistently across datasets. For the Jazz dataset, we observe that the JSD baseline significantly outperforms the majority of string-compression approaches. In contrast, for the MSD the majority of string-compression approaches significantly outperform the JSD baseline. Whereas PPM with distance $D^{\times}$ consistently outperforms all discrete-valued approaches for the Jazz dataset, PPM with compression-based NCD consistently outperforms all discrete-valued approaches for the MSD and significantly outperforms the JSD baseline.

In a comparison of continuous-valued approaches, we observe that cross-prediction using either distance $D^{\times}$ or NMSE competes with cross-correlation for the Jazz dataset. In contrast, the same cross-prediction approaches significantly outperform cross-correlation for the MSD.

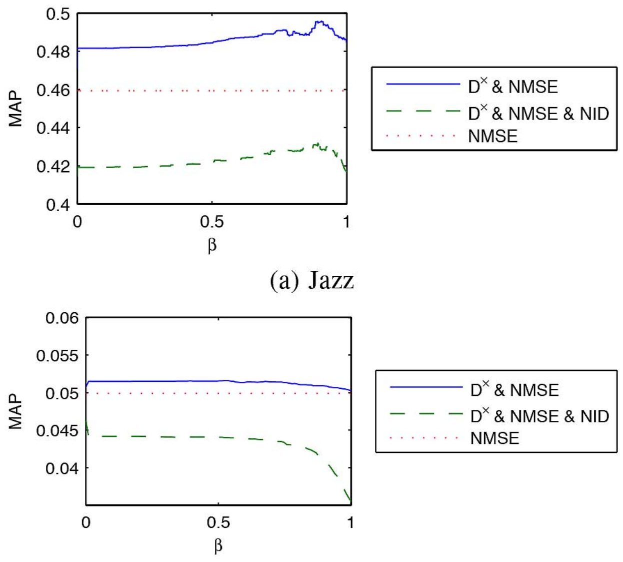
Examining continuous-valued approaches further, for both Jazz dataset and MSD, we observe a significant disadvantage in using our conditional self-prediction based estimate of NID, over cross-prediction based distances $D^{\times}$ and NMSE. The relatively poor performance of NID for the MSD suggests a limitation of our prediction approach when used with MSD chroma features. However, considering results for both datasets suggests that cross-prediction yields more favorable results than conditional self-prediction generally.

To facilitate further comparison, we consider the approaches proposed by Bertin-Mahieux and Ellis [43], Khadkevich and Omologo [57], who report MAP scores of 0.0295, 0.0371, respectively. Based on such a comparison, we obtain state-of-the-art results. Note that the stated approaches do not report any distance normalization procedure as described in Section IV-D.; we found that normalization improves our results: For the Jazz dataset and using normalized distances, we obtain MAP scores 0.425, 0.314, 0.332 for NMSE, $D^{\times}$, NID, respectively. For the MSD and using normalized distances, we obtain MAP scores 0.0340, 0.0174, 0.0216, for NMSE, $D^{\times}$, NID, respectively.

E. Combining Distances

Finally, using the method described in Section IV-G, we combine distances obtained using continuous-valued prediction. Fig. 5 displays MAP scores against mixing parameter β, for Jazz dataset and MSD. We consider the combinations $D^{\times}$&NMSE, $D^{\times}$&NMSE&NID, the latter combination which we evaluate with respect to optimal β for the former combination.

Compared to using the baseline NMSE alone, across all β and for both datasets we observe that combining NMSE with $D^{\times}$



(a) Jazz

(b) MSD

Fig. 5. Mean average precision for combinations of distances, in response to parameter β. Results displayed for Jazz dataset and MSD in subfigures (a) and (b), respectively.

TABLE IV
MEAN PRECISION AT RANK r, FOR APPROACHES BASED ON CONTINUOUS-VALUED PREDICTION

Dataset	Jazz			MSD		
r	5	10	20	5	10	20
$D^{\times}$	0.185	0.113	0.065	0.0276	0.0146	0.0077
NID	0.133	0.075	0.045	0.0147	0.0082	0.0044
NMSE	0.193	0.116	0.067	0.0270	0.0141	0.0075
$D^{\times}$ & NMSE	0.213	0.123	0.070	0.0288	0.0150	0.0079
$D^{\times}$ & NID & NMSE	0.168	0.101	0.063	0.0265	0.0146	0.0076

improves performance: For the Jazz dataset, we observe maximal MAP score 0.496, corresponding to a gain of 8.1%. For the MSD, we observe maximal MAP score 0.0516, corresponding to a gain of 3.4%. We observe no performance gain by further combining NID estimates with NMSE and $D^{\times}$, obtaining maximal MAP scores 0.432 and 0.0463 respectively for Jazz dataset and MSD. Additional evaluations revealed no performance gain using normalized distances.

Table III summarizes MAP scores; in Fig. 4(c), (d) we test for differences in performance among combinations of distances based on continuous-valued prediction. Compared to using the baseline NMSE alone, combining NMSE with $D^{\times}$ significantly improves performance for both the Jazz dataset and MSD. In addition, Table IV reports performance in terms of mean precision at ranks r. Matching previous observations, for Jazz dataset and MSD, the combination of NMSE and $D^{\times}$ consistently outperforms remaining combinations. At rank $r = 5$, relative to the NMSE baseline, we obtain a performance gain of 10.0% for the Jazz dataset and 6.7% for the MSD.

VI. CONCLUSIONS

We have evaluated measures of pairwise predictability between time series for cover song identification. We consider alternative distance measures to the NCD: We propose NCDA, which incorporates a method for obtaining joint representations of time series, in addition to methods based on cross-prediction. Secondly, we attend to the issue of representing time series: We propose continuous-valued prediction as a means of determining pairwise similarity, where we estimate compressibility as a statistic of the prediction error. We contrast methods requiring feature quantisation, against methods directly applicable to continuous-valued features.

Firstly, the proposed continuous-valued approach outperforms discrete-valued approaches and competes with evaluated continuous baseline approaches. Secondly, we draw attention to using cross-prediction as an alternative approach to the NCD, where we observe superior results in both discrete and continuous cases for Jazz cover song identification, and for the continuous case for cover song identification using the Million Song Dataset. Thirdly, using NCDA, we are able to mitigate differences in performance between evaluated discrete compression algorithms. We view the previous three points as evidence that using information-based measures of similarity, a continuous-valued representation may be preferable to discrete-valued chroma representations, owing to the challenge of obtaining discrete-valued representations. Further, NCD may yield suboptimal performance compared to alternative distance measures.

We argue that due to the ubiquity of time series similarity problems, our results are relevant to application domains extending beyond the scope of this work. Finally, in the context of cover song identification, we have demonstrated state-of-the-art performance using a large-scale dataset. We have shown that our distances based on continuous-valued prediction may be combined to improve performance relative to the baseline.

For future work, we aim to evaluate alternative time series models to those presently considered. To this end, further investigations might involve causal state space reconstruction [80] or recurrent neural networks such as the long short term memory architecture [81]. For future work, we aim to evaluate ensemble techniques for combining distances in greater detail.

REFERENCES

[1] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proc. IEEE*, vol. 96, no. 4, pp. 668–696, Apr. 2008.

[2] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of audio fingerprinting," *J. VLSI Signal Process.*, vol. 41, no. 3, pp. 271–284, 2005.

[3] J. Serrà, "Identification of versions of the same musical composition by processing audio descriptions," Ph.D. dissertation, Univ. Pompeu Fabra, Barcelona, Spain, 2011.

[4] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 638–648, Mar. 2010.

[5] M. I. Mandel and D. P. W. Ellis, "Song-level features and support vector machines for music classification," in *Proc. 6th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2005, pp. 594–599.

[6] N. Scaringella, G. Zoia, and D. J. Mlynek, "Automatic genre classification of music content: A survey," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 133–141, Mar. 2006.

[7] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 6, pp. 1138–1151, Aug. 2008.

[8] L. Meyer, *Music and Emotion*. Chicago, IL, USA: Univ. of Chicago Press, 1956.

[9] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, pp. 379, 623–423, 656, 1948.

[10] D. Huron*, Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA, USA: MIT Press, 2006.

[11] M. T. Pearce and G. A. Wiggins, “Auditory expectation: The information dynamics of music perception and cognition,” *Topics Cognitive Sci.*, vol. 4, no. 4, 2012.

[12] S. Abdallah and M. D. Plumbley, “Information dynamics: Patterns of expectation and surprise in the perception of music,” *Connect. Sci.*, vol. 21, no. 2–3, pp. 89–117, 2009.

[13] P. Foster, S. Dixon, and A. Klapuri, “Identification of cover songs using information theoretic measures of similarity,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 739–743.

[14] M. Li and P. Vitányi*, An Introduction to Kolmogorov Complexity and its Applications*. New York, NY, USA: Springer, 2008.

[15] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi, “The similarity metric,” *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3250–3264, Dec. 2004.

[16] A. Kocsor, A. Kertész-Farkas, L. Kaján, and S. Pongor, “Application of compression-based distance measures to protein sequence classification: A methodological study,” *Bioinformatics*, vol. 22, no. 4, pp. 407–412, 2006.

[17] A. Bardera, M. Feixas, I. Boada, and M. Sbert, “Image registration by compression,” *Inf. Sci.*, vol. 180, no. 7, pp. 1121–1133, 2010.

[18] S. Wehner, “Analyzing worms and network traffic using compression,” *J. Comput. Security*, vol. 15, no. 3, pp. 303–320, 2007.

[19] M. Li and R. Sleep, “Melody classification using a similarity metric based on Kolmogorov complexity,” *Sound and Music Comput.*, pp. 126–129, 2004.

[20] R. Cilibrasi, P. M. B. Vitányi, and R. Wolf, “Algorithmic clustering of music based on string compression,” *Comput. Music J.*, vol. 28, no. 4, pp. 49–67, 2004.

[21] M. Li and R. Sleep, “Genre classification via an LZ78-based string kernel,” in *Proc. 6th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2005.

[22] M. Helén and T. Virtanen, “A similarity measure for audio query by example based on perceptual coding and compression,” in *Proc. 10th Int. Conf. Digital Audio Effects (DAFX)*, 2007.

[23] T. E. Ahonen, “Measuring harmonic similarity using PPM-based compression distance,” in *Proc. Workshop Exploring Musical Inf. Spaces (WEMIS)*, 2009, pp. 52–55.

[24] J. P. Bello, “Measuring structural similarity in music,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2013–2025, Sep. 2011.

[25] T. Bertin-Mahieux, D. Ellis, B. Whitman, and P. Lamere, “The million song dataset,” in *Proc. 12th Int. Society Music Inf. Retrieval Conf. (ISMIR’11)*, Miami, FL, Oct. 24–28, 2011, pp. 591–596, Univ. of Miami, 2011.

[26] A. Berenzweig, B. Logan, D. P. W. Ellis, and B. Whitman, “A large-scale evaluation of acoustic and subjective music-similarity measures,” *Comput. Music J.*, vol. 28, no. 2, pp. 63–76, 2004.

[27] B. Logan and A. Salomon, “A music similarity function based on signal analysis,” in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, 2001, pp. 745–748.

[28] J. Aucouturier and F. Pachet, “Music similarity measures: What’s the use?,” in *Proc. 3rd Int. Conf. Music Inf. Retrieval (ISMIR)*, 2002, pp. 157–163.

[29] J. Aucouturier, B. Defreville, and F. Pachet, “The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music,” *J. Acoust. Soc. Amer.*, vol. 122, pp. 881–891, 2007.

[30] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, “Music classification via the bag-of-features approach,” *Pattern Recogn. Lett.*, vol. 32, no. 14, pp. 1768–1777, 2011.

[31] M. Helén and T. Virtanen, “Audio query by example using similarity measures between probability density functions of features,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, 2010, article 2010:179303.

[32] F. Lerdahl and R. Jackendoff*, A Generative Theory of Tonal Music*. Cambridge, MA, USA: MIT Press, 1983.

[33] M. A. Casey and M. Slaney, “The importance of sequences in musical similarity,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, vol. 5, pp. 5–8.

[34] J. Paulus, M. Müller, and A. Klapuri, “State of the art report: Audio-based music structure analysis,” in *Proc. 11th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2010, pp. 625–636.

[35] T. Fujishima, “Realtime chord recognition of musical sound: A system using common Lisp music,” in *Proc. Int. Comput. Music Conf. (ICMC)*, 1999, pp. 464–467.

[36] M. A. Bartsch and G. H. Wakefield, “To catch a chorus: Using chroma-based representations for audio thumbnailing,” in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2001, pp. 15–18.

[37] J. Foote, “ARTHUR: Retrieving orchestral music by long-term structure,” in *Proc. Int. Symp. Music Inf. Retrieval (ISMIR)*, 2000.

[38] E. Gómez and P. Herrera, “The song remains the same: Identifying versions of the same piece using tonal descriptors,” in *Proc. 7th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2006.

[39] J. Serrà, X. Serra, and R. G. Andrzejak, “Cross recurrence quantification for cover song identification,” *New J. Phys.*, vol. 11, no. 9, p. 093017, 2009.

[40] J. Serrà, H. Kantz, X. Serra, and R. G. Andrzejak, “Predictability of music descriptor time series and its application to cover song detection,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 514–525, 2012.

[41] D. P. W. Ellis and G. E. Poliner, “Identifying ‘cover songs’ with chroma features and dynamic programming beat tracking,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, vol. 4, pp. 1429–1432.

[42] J. H. Jensen, M. G. Christensen, and S. H. Jensen, “A chroma-based tempo-insensitive distance measure for cover song identification using the 2D autocorrelation function,” *Music Inf. Retrieval Eval. Exchange Task Audio Cover Song Ident.*, 2008.

[43] T. Bertin-Mahieux and D. P. W. Ellis, “Large-scale cover song recognition using the 2D Fourier transform magnitude,” in *Proc. 13th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2012, pp. 241–246.

[44] W. Tsai, H. Yu, and H. Wang, “A query-by-example technique for retrieving cover versions of popular songs with similar melodies,” in *Proc. 6th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2005, pp. 183–190.

[45] J. P. Bello, “Audio-based cover song retrieval using approximate chord sequences: Testing shifts, gaps, swaps and beats,” in *Proc. 8th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2007, pp. 239–244.

[46] K. Lee, “Identifying cover songs from audio using harmonic representation,” *Music Inf. Retrieval Eval. Exchange Task Audio Cover Song Ident.*, 2006.

[47] B. Martin, D. G. Brown, P. Hanna, and P. Ferraro, “BLAST for audio sequences alignment: A fast scalable cover identification tool,” in *Proc. 13th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2012.

[48] T. Ahonen, “Combining chroma features for cover version identification,” in *Proc. 11th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2010, pp. 165–170.

[49] J. Cleary and I. Witten, “Data compression using adaptive coding and partial string matching,” *IEEE Trans. Commun.*, vol. 32, no. 4, pp. 396–402, Apr. 1984.

[50] T. Ahonen, “Compression-based clustering of chromagram data: New method and representations,” in *Proc. 9th Int. Symp. Comput. Music Modeling Retrieval*, 2012, pp. 474–481.

[51] M. Burrows and D. J. Wheeler, A block-sorting lossless data compression algorithm, Digital Equipment Corp., Tech. Rep., 1994.

[52] I. Tabus, V. Tabus, and J. Astola, “Information theoretic methods for aligning audio signals using chromagram representations,” in *Proc. 5th Int. Symp. Commun. Control Signal Process. (ISCCSP)*, 2012, pp. 1–4.

[53] D. Silva, H. Papadopoulos, G. Batista, and D. Ellis, “A video compression-based approach to measure music structural similarity,” in *Proc. 14th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2013, pp. 95–100.

[54] M. Casey, C. Rhodes, and M. Slaney, “Analysis of minimum distances in high-dimensional musical spaces,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 1015–1028, Jul. 2008.

[55] M. Slaney and M. Casey, “Locality-sensitive hashing for finding nearest neighbors,” *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 128–131, Mar. 2008.

[56] T. Bertin-Mahieux and D. Ellis, “Large-scale cover song recognition using hashed chroma landmarks,” in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2011, pp. 117–120.

[57] M. Khadkevich and M. Omologo, “Large-scale cover song identification using chord profiles,” in *Proc. 14th Int. Society Music Inf. Retrieval Conf. (ISMIR)*, 2013, pp. 233–238.

[58] D. Schnitzer, A. Flexer, and G. Widmer, “A filter-and-refine indexing method for fast similarity search in millions of music tracks,” in *Proc. 10th Int. Society Music Inf. Retrieval Conf. (ISMIR)*, 2009, pp. 537–542.

[59] D. Loewenstern, H. Hirsh, P. Yianilos, and M. Noordewier, DNA sequence classification using compression-based induction Center for Discrete Math. Theoret. Comput. Sci., Tech. Rep., 1995.

[60] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. IT-23, no. 3, pp. 337–343, May 1977.

[61] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1270–1279, Jul. 1993.

[62] D. Benedetto, E. Caglioti, and V. Loreto, "Language trees and zipping," *Phys. Rev. Lett.*, vol. 88, no. 4, p. 48702, 2002.

[63] R. Cilibrasi and P. M. B. Vitányi, "Clustering by compression," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1523–1545, Apr. 2005.

[64] D. Sculley and C. E. Brodley, "Compression and machine learning: A new perspective on feature space vectors," in *Proc. Data Compression Conf. (DCC)*, 2006, pp. 332–341.

[65] P. Grünwald and P. M. B. Vitányi, "Shannon information and Kolmogorov complexity," *arXiv e-print cs/0410002*, 2004.

[66] A. Kaltchenko, "Algorithms for estimating information distance with application to bioinformatics and linguistics," in *Proc. IEEE Can. Conf. Elect. Comput. Eng. (CCECE)*, 2004, vol. 4, pp. 2255–2258.

[67] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inf. Theory*, vol. 38, no. 4, pp. 1258–1270, Jul. 1992.

[68] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 259–266, Jan. 1994.

[69] R. Begleiter, R. El-Yaniv, and G. Yona, "On prediction using variable order Markov models," *J. Artif. Intell. Res.*, vol. 22, pp. 385–421, 2004.

[70] F. Takens, "Detecting strange attractors in turbulence," *Dynam. Syst. Turbul.*, pp. 366–381, 1981.

[71] E. Gómez, "Tonal description of music audio signals," Ph.D. dissertation, Univ. Pompeu Fabra, Barcelona, Spain, 2006.

[72] P. Foster, A. Klapuri, and M. D. Plumbley, "Causal prediction of continuous-valued music features," in *Proc. 12th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2011, pp. 501–506.

[73] D. P. W. Ellis, "Beat tracking with dynamic programming," *J. New Music Res.*, vol. 36, no. 1, pp. 51–60, 2007, Taylor & Francis.

[74] T. Jehan, Analyzer documentation The Echo Nest, Tech. Rep., 2011.

[75] S. Ravuri and D. P. W. Ellis, "Cover song detection: From high scores to general classification," in *Proc. IEEE Int. Conf. Acous. Speech Signal Process. (ICASSP)*, 2010, pp. 65–68.

[76] J. Osmalskyj, S. Pierard, M. Van Droogenbroeck, and J. Embrechts, "Efficient database pruning for large-scale cover song recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 714–718.

[77] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Statist. Assoc.*, vol. 32, no. 200, pp. 675–701, 1937.

[78] J. W. Tukey, *The Problem of Multiple Comparisons*. Princeton, NJ, USA: Princeton Univ., 1973.

[79] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[80] C. R. Shalizi and K. L. Shalizi, "Blind construction of optimal nonlinear recursive predictors for discrete sequences," in *Proc. 20th Conf. Uncertainty Artif. Intell. (UAI)*, 2004, pp. 504–511.

[81] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

Peter Foster is a Ph.D. candidate at the Centre for Digital Music, Queen Mary University of London. He received the M.Sc. and B.Sc. degrees in computer science from the University of Edinburgh and from the University of East Anglia, respectively. He has worked as a Software Developer in the public sector.

At the Centre for Digital Music, his research interests include music content analysis, musical similarity, time series modeling, machine learning, and information theory.

Simon Dixon has a Ph.D. in computer science (Sydney) and LMusA diploma in classical guitar. He is a Reader (Assoc. Prof.) at Queen Mary University of London, where he leads research on music informatics in the Centre for Digital Music. His research interests include high-level music signal analysis and the representation of musical knowledge (particularly rhythm and harmony), and has published over 100 refereed papers in the area of music informatics. He is President of the International Society for Music Information Retrieval.

Anssi Klapuri received his Ph.D. degree from the Tampere University of Technology (TUT), Tampere, Finland, in 2004. He visited as a Post-Doc Researcher at Ecole Centrale de Lille, France, and Cambridge University, U.K., in 2005 and 2006, respectively. He worked as a Lecturer at the Centre for Digital Music at Queen Mary University of London, London, U.K., in 2010–2011. He is currently CTO at Yousician, Finland, and Adjunct Professor at TUT. His research interests include audio signal processing, auditory modeling, and machine learning.