IDENTIFICATION OF COVER SONGS USING INFORMATION THEORETIC MEASURES OF SIMILARITY

Peter Foster, Simon Dixon

Centre for Digital Music
Queen Mary University of London
London, United Kingdom
{peter.foster,simon.dixon}@eecs.qmul.ac.uk

Anssi Klapuri

Ovelin
Helsinki, Finland
anssi@ovelin.com

ABSTRACT

We consider techniques for cover song detection, based on information theoretic notions of compressibility. We propose methods for computing the normalised compression distance (NCD), while accounting for correlation between time series. Secondly, we describe methods based on cross-prediction for estimating compressibility between sequences of continuous-valued features. Using the latter approach, we view the NCD as a statistic of the prediction error. We evaluate the proposed approaches using a data set consisting of 300 Jazz songs. Quantified in terms of mean average precision, the proposed continuous-valued approach outperforms considered quantisation-based approaches.

Index Terms— Cover song detection, normalised compression distance, audio similarity measures, time series prediction

1. INTRODUCTION

We consider the problem of *cover song detection*, where a cover song is defined as a rendition of a piece of music [1]. Thus, we do not require that renditions of a piece of music are produced by the same artist. Given a piece of musical audio presented as a query, the objective of cover song detection is to identify tracks in a data set which may be considered renditions of the query. Potential applications of cover song detection include automated music recommendation, copyright infringement detection, and musicological research [1].

An important task in cover song detection involves quantifying pairwise musical similarity between tracks, to determine whether said tracks should be considered cover versions of one another. Note that cover versions may be subject to alterations in instrumentation, tempo, expressive performance, melody, harmony, lyrics, and musical form. Quantifying similarity for cover song detection therefore presents a challenging task. Correspondingly, a variety of methods have been proposed for representing and modelling music signals for cover song detection, involving cross-correlation [2], alignment techniques [3], cross-recurrence analysis [4, 5], and time series prediction [6].

In this work, we propose novel approaches for determining musical similarity, based on statistical modelling of audio feature time series. In particular, we adopt an information theoretic approach to quantifying joint compressibility between time series [7], the latter of which has recently been applied in the context of cover song detection [5, 8, 9]. As described in Section 3.1, we propose methods for jointly aggregating pairs of time series, for subsequent modelling using compressibility as a measure of musical similarity. Furthermore, as described in Section 3.2, we propose a novel approach based on modelling continuous-valued features directly, without requiring quantisation.

2. BACKGROUND

Let us denote with $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_N), \mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_M)$ a pair of time series, each corresponding to a sequence of continuous-valued feature vectors extracted from musical audio. In this work, we assume that $\mathbf{V}, \mathbf{U}$ have been generated by a pair of information sources X, Y, respectively.

If we assume that both sources X, Y generate sequences of independent and identically distributed observations with respective probability densities $p_X(\cdot), p_Y(\cdot)$, one possible measure of dissimilarity between time series is the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(p_X \| p_Y)$, defined as

$$D_{\mathrm{KL}}(p_X \| p_Y) = \int \log \left(\frac{p_X(\mathbf{x})}{p_Y(\mathbf{x})} \right) p_X(x) \, \mathrm{d}\mathbf{x}. \tag{1}$$

The KL divergence has been applied widely in music content analysis [10] as a 'bag of features' approach, referring to the notion that temporal order between features is discarded. Recall that taking the logarithm in Equation 1 to base 2, $D_{\mathrm{KL}}(p_X \| p_Y)$ quantifies the expected number of additional bits required to encode observations from source X, given a code for source Y.

To account for temporal order in musical audio, the normalised compression distance (NCD) [7] has recently been applied to sequences of quantised music features [8, 9]. The NCD between two strings $x = (x_1, x_2, \ldots, x_N), y = (y_1, y_2, \ldots, y_M)$ is defined as

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

where $C(\cdot)$ denotes the compressed size of an input string in bits and where xy denotes the concatenation of x and y. By applying a compressor such as the Lempel-Ziv (LZ) algorithm [11], the NCD may be taken as an approximation of the normalised information distance (NID) [12], defined as

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

where the uncomputable function $K(\cdot)$ denotes the binary length of the shortest program which outputs the given string [12]. Similarly, $K(x, y)$ denotes the binary length of the shortest program which outputs both x, y, in addition to delimiting and specifying the order of output strings. Assuming the identity

$$K(x|y) = K(x, y) - K(y) \tag{4}$$

where $K(x|y)$ denotes the binary length of the shortest program which outputs x, given input string y, and assuming that $K(x,y) = K(y,x)$, Equation 3 may be rewritten as

$$\mathrm{NID}(x,y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}. \tag{5}$$

As given in Equation 5, the numerator may be interpreted as quantifying the amount of information disparity between x, y, whereas the denominator ensures that values lie within the unit interval. The NID is optimal within the class of all metrics, in the sense that it incorporates the most relevant feature for comparing strings [12]. Since the NCD is conceived as a computable approximation to the NID, the choice of compression approach may be considered to define a feature space used to quantify similarity [12].

3. METHOD

3.1. Sequence representation and the NCD

Note that the term $C(xy)$ in Equation 2 proscribes that strings are concatenated, as an approximation to $K(x,y)$. Assuming finite-order Markov sources X, Y with stationary transition probabilities, 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

$$H_\mu(X) = \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} \tag{6}$$

$$H_\mu(X,Y) = \lim_{n\to\infty} \frac{H((X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n))}{n} \tag{7}$$

$$H_\mu(X|Y) = H(X,Y) - H(Y). \tag{8}$$

Following [13], we approximate the quantities $K(x)$, $K(xy)$, $K(x|y)$ as

$$K(x) \approx H_\mu(X)\,|x| \tag{9}$$

$$K(xy) \approx H_\mu(X,Y)\,|x| \tag{10}$$

$$K(x|y) \approx H_\mu(X|Y)\,|x| \tag{11}$$

where $|x|$ denotes the length of x and where we assume that realisations of X, Y are given by x, y, respectively. As observed in [13], Equation 7 accounts for correlation between sources, thus source coding should be performed on the sequence of pairs $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$. Based on this observation, we propose the normalised compression distance with alignment (NCDA), defined as

$$\mathrm{NCDA}(x,y) = \frac{C(\langle x,y\rangle) - \min\{C(x), C(y))\}}{\max\{C(x), C(y)\}} \tag{12}$$

where $\langle x,y\rangle$ instead of concatenating inputs, aligns x, y as a means of maximising temporal correlation. In this work, we proceed by equalising time series lengths, before determining the rotation of y along the time axis which maximises cross-correlation at zero lag between time series. As proposed in [13], we then interleave x, y, generating a string of super-symbols from the cross alphabet $\mathcal{X}\times\mathcal{Y}$, where $\mathcal{X}$, $\mathcal{Y}$ denote the alphabets of strings x, y, respectively.

3.2. Continuous-valued prediction and the NCD

One possible approach to using the NCD involves coding discrete-valued time series, using general purpose compression techniques. Applied to audio features, this approach requires addressing the non-trivial task of determining a suitable quantisation scheme [8, 9].

In this work, we propose an alternative approach based on coding unquantised audio features. We assume two series of k-dimensional real-valued feature vectors $\mathbf{V}, \mathbf{U}$. For a strongly stationary source, the entropy rate $H_\mu(X)$ is equivalent to

$$H_\mu(X) = \lim_{m\to\infty} H(X_m | X_{m-1}, \ldots, X_1) \tag{13}$$

which may be interpreted as the uncertainty about observation X_m, given preceding observations $X_1,\ldots,X_{m-1}$, in the limit as $m\to\infty$. We consider the aforementioned interpretation in terms of prediction. Thus, we define with $\tilde{\mathbf{v}}_{n+1}$ the successor of $\mathbf{v}_1,\ldots,\mathbf{v}_n$, as forecast by a continuous predictor $\mathcal{F}$ using some model of observations $\mathcal{M}(\mathbf{V})$,

$$\tilde{\mathbf{v}}_{n+1} = \mathcal{F}\left(\mathbf{v}_1,\ldots,\mathbf{v}_n,\, \mathcal{M}(\mathbf{V})\right). \tag{14}$$

We denote with $\tilde{\mathbf{V}} = (\tilde{\mathbf{v}}_1,\ldots,\tilde{\mathbf{v}}_N)$ the sequence of predictions obtained using $\mathcal{F}$ under model $\mathcal{M}(\mathbf{V})$, and denote with $v_{i,n}$ the ith component of $\mathbf{v}_n$. Furthermore, we denote with $\boldsymbol{\epsilon}_n$ the rescaled error between $\tilde{\mathbf{v}}_n$ and $\mathbf{v}_n$, whose ith component is given by

$$\epsilon_{i,n} = \frac{\tilde{v}_{i,n} - v_{i,n}}{s_i} \tag{15}$$

where s_i corresponds to the biased sample variance of the ith component in $\mathbf{V}$. Assuming an independent random variable Z, whose samples are given as $Z_1 = \boldsymbol{\epsilon}_1,\ldots,Z_N = \boldsymbol{\epsilon}_N$, we estimate $H_\mu(X)$ as the entropy of the prediction error, $H(Z)$. We then estimate $K(x)$ using Equation 9, proceeding analogously for $K(y)$.

Viewed in terms of continuous-valued prediction, we use the definition of NID given in Equation 5 and estimate $H_\mu(X|Y)$ as the entropy of the prediction error, with the amendment that prediction $\tilde{\mathbf{v}}'_{n+1}$ is instead forecast by $\mathcal{F}$ using the model $\mathcal{M}(\mathbf{V},\mathbf{U})$,

$$\tilde{\mathbf{v}}'_{n+1} = \mathcal{F}\left(\mathbf{v}_1,\ldots,\mathbf{v}_n,\, \mathcal{M}(\mathbf{V},\mathbf{U})\right). \tag{16}$$

Equation 16 corresponds to cross-prediction, where observations $\mathbf{V}$ are predicted causally and where prior observations in both $\mathbf{V}$, $\mathbf{U}$ inform predictions. We then estimate $K(x|y)$ using Equation 11, proceeding analogously for $K(y|x)$.

3.2.1. Prediction method

To form predictions $\tilde{\mathbf{V}}$, as proposed in [4, 5], we adopt a dynamical systems approach and construct a time delay embedding [14], whose elements $\mathbf{s}^{\mathbf{c}}_r$ constructed from a time series $\mathbf{C}$ are given by

$$\begin{aligned}\mathbf{s}^{\mathbf{c}}_r = (&c_{1,r}, c_{1,(r-1)\tau}, \ldots, c_{1,(r-d+1)\tau},\\ &\ldots, c_{k,r}, c_{k,(r-1)\tau}, \ldots, c_{k,(r-d+1)\tau}).\end{aligned} \tag{17}$$

In Equation 17, the amount of temporal context accounted for in state vector $\mathbf{s}^{\mathbf{c}}_r$ is controlled by the embedding dimension d and time delay τ, with $d \le r$ and $\tau > 0$. Given observation context $\mathbf{v}_{n-d+1},\ldots,\mathbf{v}_n$, we predict $\tilde{\mathbf{v}}_{n+h}$ under $\mathcal{M}(\mathbf{V},\mathbf{U})$ using a nearest-neighbour approach,

$$\tilde{\mathbf{v}}_{n+h} = \mathbf{w}_{q+h}, \quad q = \underset{r\in[d\,..\,M-h]\setminus\{n\}}{\arg\max} \mathrm{corr}(\mathbf{s}^{\mathbf{w}}_r, \mathbf{s}^{\mathbf{v}}_n). \tag{18}$$

where $\mathrm{corr}(\cdot,\cdot)$ denotes Pearson's correlation coefficient, where h specifies the prediction horizon and where $\mathbf{w}_1,\ldots,\mathbf{w}_{N+M}$ represents the concatenated time series $[\mathbf{VU}]$. Note that we disregard the trivial prediction $\tilde{\mathbf{v}}_{n+h} = \mathbf{v}_{n+h}$. We perform prediction under $\mathcal{M}(\mathbf{V})$ analogously. Assuming normally distributed prediction errors Z with covariance $\boldsymbol{\Sigma}$, we compute $H(Z)$ as

$$H(Z) = \frac{1}{2}\log\left(2\pi e\right)^k |\boldsymbol{\Sigma}| \tag{19}$$

where we estimate $\mathbf{\Sigma}$ from the sample covariance and where k denotes the dimensionality of features. In this approach, the NCD is a function of statistics of the prediction error sequence. Thus, the NCD may be contrasted with the cross-prediction mean squared error (MSE), the latter which may be applied as a measure of dissimilarity assuming independent error components [6]. As an alternative to Equation 5, we compute the distance D_+, given as

$$D_+(X,Y) = \frac{H'_\mu(X|Y) + H'_\mu(Y|X)}{H_\mu(X) + H_\mu(Y)} \tag{20}$$

where we estimate $H'_\mu(X|Y)$ using cross-prediction based on sequence $\mathbf{V}$ alone, instead of the concatenated time series $[\mathbf{VU}]$ used to estimate $H_\mu(X|Y)$.

4. EVALUATION

We evaluate the proposed approach using $L = 300$ recordings of Jazz standards. Tracks are identified as cover songs based on their title strings. Henceforth, we refer to audio tracks using the set $A = [1\,..\,L]$. Title string identities define a partition $\mathcal{P}$ on A. We use the set-valued map $\mathcal{C} : A \rightarrow \mathcal{P}$ to determine the equivalence class containing tracks deemed to be covers of the jth track, with $1 \leq j \leq L$. The entire data set has an average equivalence class size of 3.06 tracks, with a minimum equivalence class size of 2 tracks.

4.1. Feature extraction

As a descriptor of musical harmonic content, we extract 12-component beat averaged chroma features, using the method described in [2]. In this approach, chroma extraction is based on mapping FFT bins to pitch classes in the chromatic scale, assuming equal-tempered tuning and adjusting for global tuning deviations of up to 0.5 semitones. Chroma vectors are scaled with respect to the Euclidean norm. A beat tracking stage first achieves onset detection by computing first-order differences along the time scale of a log-magnitude Mel-frequency spectrogram. Next, a global tempo estimate is computed by determining maxima in the autocorrelated onset signal, where the autocorrelation signal itself is windowed as a means of specifying a preferred beat rate B. Finally, beats are determined by optimising with respect to onset magnitudes and estimated global tempo, using dynamic programming.

We account for key variation between pairs of tracks by computing the optimal transposition index (OTI) [3]. In the latter approach, we represent global harmonic content vectors $\mathbf{h_U}, \mathbf{h_V}$ by computing the averages of chroma sequences $\mathbf{U}, \mathbf{V}$. We then rotate feature vectors in $\mathbf{V}$ by the amount which maximises the inner product between $\mathbf{h_U}$ and $\mathbf{h_V}$.

4.2. Quantisation

To quantise feature vectors, we apply k-means clustering, with codebook sizes in the range $[2\,..\,48]$. To improve stability, we quantise 20 times and select the clustering which minimises the mean squared error between observations and assigned centroids. Given a pair of time series $(\mathbf{V}, \mathbf{U})$, we proceed by constructing a codebook on the first time series $\mathbf{V}$ before assigning observations in both $\mathbf{V}, \mathbf{U}$ to clusters, using the obtained codebook. Since the order of $\mathbf{V}, \mathbf{U}$ affects the outcome of clustering, we average pairwise distances computed for both possible orderings.

4.3. Distance measures

Since the KL divergence as defined in Equation 1 is non-symmetric, we apply the additional step of computing the Jensen-Shannon (JS) divergence $D_{\mathrm{JS}}(p_X \| p_Y)$, defined as

$$D_{\mathrm{JS}}(p_X \| p_Y) = D_{\mathrm{KL}}(p_X \| p_A) + D_{\mathrm{KL}}(p_Y \| p_A) \tag{21}$$

where p_A is given by the mean of p_X, p_Y,

$$p_A = \frac{1}{2}\left(p_X + p_Y\right). \tag{22}$$

We compute the JS divergence between normalised histograms of symbols, obtained by quantising pairs of time series.

We compute the discrete NCD and NCDA in combination with the following algorithms: Prediction by partial matching (PPM) [15], Burrows-Wheeler (BW) compression [16] and LZ compression [11], implemented respectively as PPMD, BZIP2 and ZLIB in the CompLearn toolkit[1], where we set parameters to favour compression rates over computational cost. In the case of NCDA, we equalise string lengths by padding the shorter of two strings, using the mode of observations. For the continuous NCD, we evaluate using parameters $h \in \{1, 4, 6\}$, $d \in \{4\}$, $\tau \in \{1, 2\}$.

For all evaluated approaches, we normalise pairwise distances between tracks using the method described in [17], as a means of compensating for cover song candidates consistently deemed similar to query tracks.

4.4. Performance statistics

Following [5], we evaluate performance using mean average precision (MAP). For a given query track j, we rank the remaining $L-1$ tracks by ascending distance and denote with $\mathcal{R}(r, j)$, the track at rank r. We define with $\Omega(r, j)$ an indicator of the relevance of track $\mathcal{R}(r, j)$,

$$\Omega(r,j) = \begin{cases} 1 & \text{if } \mathcal{R}(r,j) \in \mathcal{C}(j) \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

and define with $P_j(r)$ the precision at rank r,

$$P_j(r) = \frac{1}{r}\sum_{c=1}^{r} \Omega(c,j). \tag{24}$$

The average precision AP_j for the jth query is then defined as

$$\mathrm{AP}_j = \frac{1}{|\mathcal{C}(j)|}\sum_{r=1}^{L} P_j(r)\Omega(r,j). \tag{25}$$

We average AP_j over all L queries to obtain the MAP. Following [5], we test for statistical significance using Friedman's test [18], to account for non-normally distributed AP_j. As displayed in Table 1, the Friedman test is based on ranking each considered approach by its AP_j value, for given j. We then average ranks over all L queries, obtaining the mean rank for a given approach. We apply Tukey's range test [19] to facilitate multiple comparisons.

5. RESULTS

Figure 1 displays the effect of preferred beat rate B and codebook size on MAP, using the LZ compressor with NCDA. As observed,

[1] http://www.complearn.org/download.html

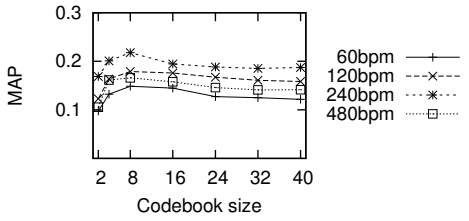
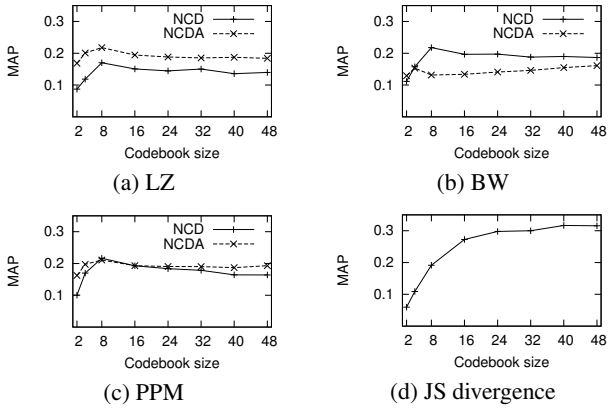


Fig. 1. Effect of B on MAP, using LZ compressor and NCDA. See main text for a description of approaches.



(a) LZ (b) BW

(c) PPM (d) JS divergence

Fig. 2. Effect of codebook size on mean average precision, using discrete-valued approaches. Results obtained for $B = 240$ bpm.

for all considered codebook sizes, the MAP peaks at $B = 240$ bpm. We hence report following results using the latter feature resolution.

Figures 2 (a)–(d) display results obtained using discrete-valued approaches, where we consider the effect of codebook size on MAP, using both NCD and NCDA approaches in Figures (a)–(c).

In combination with PPM and LZ compressors, with the exception of codebook size 8 and PPM, NCDA consistently competes with NCD, in the case of LZ compression by an average margin of 42%. However, for BW compression the relation between NCD and NCDA is reversed. A possible cause is that BW is not stream based and thus does not assume Markov sources [20]. Therefore, in the latter case we expect no improvement using NCDA in place of NCD.

In Table 1, we compare approaches, where MAP performance has been maximised over respective parameter spaces. At the 95% level, the proposed continuous approach combined with distance D_+ outperforms discrete-valued and continuous-valued NCD. Among discrete-valued approaches, we find no significant effect of interchanging compressors. However, in the case of LZ compression, utilising NCDA significantly improves performance.

Compared to a baseline using the JS divergence, we obtain significantly lower MAP scores using discrete-valued NCD approaches. In further comparison, we observe that the continuous NCD distance based on Equation 5 is outperformed by the aforementioned baseline. However, distance D_+ yields competitive performance, both relative to the cross-correlation approach described in [2], and relative to cross-prediction using the mean squared error (MSE).

6. CONCLUSIONS AND FURTHER WORK

We have considered the problem of cover song detection, where we utilise measures based on the NCD to determine similarity between pairs of audio feature time series. To this end, we have proposed a method for computing the joint compressibility of two quantised time series, while accounting for correlation between the time series, which we refer to as NCDA. As an alternative approach, we have proposed methods based on the NCD applicable to unquantised time series. Using the latter approach, our information measures may be interpreted as statistics of the cross-prediction error.

Evaluated against a collection of Jazz songs, results suggest that NCDA may bear relevance in combination with PPM and LZ compressors. Furthermore, we observe that the proposed approach based on continuous cross-prediction outperforms discrete-valued NCD and NCDA.

Considering that cover song detection performance is significantly affected by the choice of distance measure, we aim to examine in greater detail the performance and properties of alternative information-based similarity measures. Furthermore, we aim to evaluate the effect of utilised features and quantisation methods. In addition, we aim to evaluate combinations of pairwise distance measures, following the approach described in [17].

Method	NCDA		NCD	
	MAP	Mean rank	MAP	Mean rank
PPM	0.211	5.75	0.216	6.36
BW	0.161	4.69	0.218	6.26
LZ	0.218	6.00	0.170	4.81
	MAP		Mean rank	
Continuous D_+	0.431		9.12	
Continuous NCD	0.228		5.56	
JS divergence	0.317		7.66	
MSE	0.452		9.43	
Ellis and Poliner [2]	0.436		8.92	
Random	0.028		3.45	

Table 1. Summary of MAP results for evaluated approaches. The confidence interval obtained during post-hoc analysis with $\alpha = .05$ is $\pm.461$, with respect to each mean rank value. 'Random' denotes sampling pairwise distances from a normal distribution.

7. RELATION TO PRIOR WORK

This work has considered the normalised compression distance as a measure of musical similarity, for audio based cover song detection. Whereas existing studies examine the effect of audio features on performance, in combination with general purpose compression algorithms [8, 5, 9], this work examines the type of feature representation required to compute joint compressibility in the NCD. Our approach is based on the observation that the NCD does not account for correlation between sources [13]. To our knowledge, this approach has to date not been considered in the wider literature, in which the NCD has been utilised extensively [7, 5]. Furthermore, whereas cross-predictability measures have been proposed to determine similarity between continuous-valued time series [6], to our knowledge an approach based on the NCD has not been considered to date.

8. ACKNOWLEDGEMENTS

This work was supported by funding from the United Kingdom Engineering and Physical Sciences Research Council (EPSRC). We thank the anonymous reviewers for their comments. We thank Prof. Dan Ellis for making available the implementation of the method described in [2]. Finally, we thank Dr. Rudi Cilibrasi for making available the CompLearn toolkit software.

9. REFERENCES

[1] J. Serrà, *Identification of versions of the same musical composition by processing audio descriptions*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, 2011.

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

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

[4] J. Serrà, X. Serra, and R.G. Andrzejak, "Cross recurrence quantification for cover song identification," *New Journal of Physics*, vol. 11, no. 9, 2009.

[5] J.P. Bello, "Measuring structural similarity in music," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2013–2025, 2011.

[6] 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, and Language Processing*, vol. 20, no. 2, pp. 514–525, 2012.

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

[8] T.E. Ahonen, "Measuring harmonic similarity using ppm-based compression distance," in *Proc. Workshop on Exploring Musical Information Spaces*, 2009.

[9] I. Tabus, V. Tabus, and J. Astola, "Information theoretic methods for aligning audio signals using chromagram representations," in *Proc. 5th Intern. Symposium on Communications Control and Signal Processing (ISCCSP)*. IEEE, 2012, pp. 1–4.

[10] A. Berenzweig, B. Logan, D.P.W. Ellis, and B. Whitman, "A large-scale evaluation of acoustic and subjective music-similarity measures," *Computer Music Journal*, vol. 28, no. 2, pp. 63–76, 2004.

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

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

[13] A. Kaltchenko, "Algorithms for estimating information distance with application to bioinformatics and linguistics," in *Proc. Canadian Conference on Electrical and Computer Engineering*. IEEE, 2004, vol. 4, pp. 2255–2258.

[14] F. Takens, "Detecting strange attractors in turbulence," *Dynamical systems and turbulence*, pp. 366–381, 1981.

[15] J. Cleary and I. Witten, "Data compression using adaptive coding and partial string matching," *IEEE Trans. Communications*, vol. 32, no. 4, pp. 396–402, 1984.

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

[17] S. Ravuri and D.P.W. Ellis, "Cover song detection: from high scores to general classification," in *Proc. Intern. Conf. Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 65–68.

[18] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.

[19] J.W. Tukey, *The problem of multiple comparisons*, Princeton University, 1973.

[20] H. Kaplan and E. Verbin, "Most Burrows-Wheeler based compressors are not optimal," in *Combinatorial Pattern Matching*. Springer, 2007, pp. 107–118.