

Sequential Complexity as a Descriptor for Musical Similarity

Peter Foster, *Student Member, IEEE*, Matthias Mauch, *Member, IEEE* and Simon Dixon

Abstract—We propose string compressibility as a descriptor of temporal structure in audio, for the purpose of determining musical similarity. Our descriptors are based on computing track-wise compression rates of quantised audio features, using multiple temporal resolutions and quantisation granularities. To verify that our descriptors capture musically relevant information, we incorporate our descriptors into similarity rating prediction and song year prediction tasks. We base our evaluation on a dataset of 15 500 track excerpts of Western popular music, for which we obtain 7 800 web-sourced pairwise similarity ratings. To assess the agreement among similarity ratings, we perform an evaluation under controlled conditions, obtaining a rank correlation of 0.33 between intersected sets of ratings. Combined with bag-of-features descriptors, we obtain performance gains of 31.1% and 10.9% for similarity rating prediction and song year prediction. For both tasks, analysis of selected descriptors reveals that representing features at multiple time scales benefits prediction accuracy.

Index Terms—Music content analysis, musical similarity measures, time series complexity

I. INTRODUCTION

We are concerned with the task of quantifying musical similarity, which has received considerable interest in the field of audio-based music content analysis [1], [2]. Owing to the proliferation of music in digital formats and the expansion of web-based music databases, there is an impetus to develop novel search, navigation and recommendation systems. Music content analysis has found application in such information retrieval systems as an alternative to manual annotation processes, when the latter are infeasible, unavailable or amenable to be supplemented [3].

We may distinguish between music content analysis applications such as audio fingerprinting [4], version identification [5], genre classification [6] and mood identification [7]. Given a query track, audio fingerprinting typically should identify a unique track deemed similar with respect to a collection. In contrast, for genre and mood classification, the set of tracks deemed similar with respect to a collection is typically large. Thus, we may distinguish between music classification tasks according to the degree of *specificity* associated with the measure of musical similarity [1].

Copyright © 2014 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

P.F. is funded by an Engineering and Physical Sciences Research Council Doctoral Training Account studentship. M.M. is funded by a Royal Academy of Engineering Research Fellowship.

All authors are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, UK. (Email: peter.foster@eecs.qmul.ac.uk; matthias.mauch@eecs.qmul.ac.uk; simon.dixon@eecs.qmul.ac.uk)

In this work, we consider two low-specificity tasks, namely similarity rating prediction and song year prediction. An important issue in our considered domain surrounds feature representation. In particular, we address the problem of representing temporal structure in audio features. We refer to summary statistics of audio features extracted from a song as descriptors. Descriptors may be characterised according to how temporal structure is accounted for [2]. We may distinguish between *bag-of-features* representations [8], which discard information on temporal structure, and sequential representations. As a sequential representation, we propose to estimate the complexity of audio feature time series, where we quantify complexity in terms of string compressibility. As a result, we obtain scalar-valued summary statistics which retain information on temporal structure.

We motivate our evaluations involving similarity rating prediction and song year prediction to test the hypothesis that our complexity descriptors capture temporal information in audio features and that such information is relevant for determining musical similarity. For similarity rating prediction, our ground truth is given by human similarity judgements and we assume that an objective musical similarity correlates with subjects' degree of perceived musical similarity, based on a five-point rating scale. For song year prediction, our ground truth is readily given by chart entry times of songs and we assume that musical similarity correlates with chart entry time proximity. Whereas song year prediction has received little attention in the literature, the song year is important in determining musical preference [9]. Thus, song year prediction might be applied in music recommendation [10]. Song year prediction might furthermore be incorporated in genre classification tasks, since musical genres are associated with particular years.

Section II provides an overview of methods and descriptors for computing low-specificity similarity. In Section III, we describe our approach. In Section IV, we detail our experimental method and results; we provide separate accounts for similarity rating prediction and song year prediction in Sections IV-A and IV-B, respectively. Finally, in Section V we provide conclusions.

II. BACKGROUND

For a detailed review of recent literature on methods for determining musical similarity, from the perspective of classification, we refer to the work of Fu et al. [2]. To determine musical similarity, one possible approach involves computing pairwise distances between tracks. The obtained distances may then be used for classification. A second approach consists in applying track-wise descriptors directly for classification.

Based on the second approach, Tzanetakis and Cook [11] compute first and second-order moments on spectral features including MFCCs, to perform genre classification using the k -nearest neighbours (KNN) algorithm and Gaussian mixture models (GMMs) estimated on each target class. Li and Ogi-hara [12] propose to classify Daubechies wavelet histograms using GMMs and KNN for genre and mood classification. Using spectral features, West et al. [13] propose methods for learning similarity functions based on constructing decision trees for genre classification. Slaney et al. [14] propose feature transformations based on supervised learning and using onset and loudness features, for the purpose of album and artist classification.

Based on the approach of determining distances between descriptors, Logan and Salomon [15] propose to estimate GMMs on individual tracks. Pairwise track distances are then computed using a combination of Kullback-Leibler divergence (KLD) and earth mover's distance, where the KLD is used to compare pairs of track centroids. The approach based on KLD assumes that each centroid follows a Gaussian distribution; thus the KLD may be computed in closed form as

$$\text{KLD} = \frac{1}{2} \left(\text{tr}(\Sigma_1^{-1}\Sigma_2) + (\mu_1 - \mu_2)^T \Sigma_1^{-1} (\mu_1 - \mu_2) - h - \log \frac{|\Sigma_2|}{|\Sigma_1|} \right) \quad (1)$$

where Σ_1, Σ_2 and μ_1, μ_2 respectively denote the mean and covariance of two multivariate Gaussian distributions with dimensionality h . Aucouturier and Pachet [16] in contrast compute cross-likelihoods between GMMs using Monte Carlo approximations for the purpose of genre classification, whereas Berenzweig et al. [17] consider the asymptotic likelihood approximation of the KLD and centroid distances for the task of similarity rating prediction. Mandel and Ellis [18] instead represent tracks as single Gaussians and use (1) as a distance measure between track pairs. The obtained distances are then applied to artist identification, using support vector machines (SVMs) for classification. An alternative approach to computing the KLD is based on computing histograms of quantised features, as proposed by Vignoli and Pauws [19] for playlist recommendation; Levy and Sandler [20] compare approaches in the context of genre classification.

The previously described techniques are commonly referred to *bag-of-features* approaches, since they discard information on temporal structure. Yet, the relative convenience of bag-of-features approaches stands in contrast to the importance of temporal structure in perception of musical timbre, as observed by McAdams et al. [21]. Aucouturier and Pachet [8] argue that the bag-of-features approach is insufficient to model polyphonic music for determining similarity. Sequential representations based on mid-level features are widely applied for the purpose of version identification [5]. For low-specificity classification, one possible approach to mitigating the shortcoming of the *bag-of-features* approach involves the intermediate step of aggregating features locally, before summarising anew using obtained summary statistics. Tzanetakis and Cook [11] propose to estimate the local mean and variance of features contained in a 1s window. For the task of predicting

musical similarity, Seyerlehner et al. [22] apply a single, global summarisation step to overlapping windows, computing variance and percentiles. For the purpose of local aggregation, alternative pooling functions are considered by Mörchen et al. [23], Hamel et al. [24], Wülfing and Riedmiller [25].

An alternative approach is based on retaining the temporal order of features at each window position. Spectral analysis may be applied to the original features, resulting in a new feature sequence. Pampalk [26] proposes fluctuation patterns describing loudness modulation across frequency bands, whereas Lee et al. [27] propose statistics based on modulation spectral analysis. Mörchen et al. [23] consider a variety of statistics based on spectral analysis and autocorrelation. Meng et al. [28], Coviello et al. [29] apply multivariate autoregressive modelling to windowed features, for the tasks of genre and tag classification.

To account for temporal structure, statistical modelling may be applied to quantised features. For genre classification, Li and Sleep [30] propose an SVM kernel in which pairwise distances are obtained by comparing dictionaries generated using the Lempel-Ziv compression algorithm [31]. Reed and Lee [32] apply latent semantic analysis to unigram and bigram counts for classification using SVMs, whereas Langlois and Marques [33] propose to estimate language models for computing sequence cross-likelihoods for genre and artist classification. Ren and Jang [34] propose an algorithm for computing histograms of feature codeword sequences for genre classification.

Recent approaches attempt to model temporal structure using representations constructed at multiple time scales. Based on a bag-of-features approach, Foucard et al. [35] propose an ensemble of classifiers, where each classifier is trained on features at a given time scale. Features at successive resolutions are aggregated using averaging. Applied to tag and instrument classification, results indicate that a multiscale approach benefits performance. Dieleman and Schrauwen [36] apply feature learning based on spherical K -means clustering to tag classification. Evaluated aggregation techniques are based on varying the spectrogram window size, in addition to Gaussian and Laplacian pyramid smoothing techniques. Although not applied to classification, Mauch and Levy [37] propose a similar smoothing approach for characterising structural change at multiple time scales. Finally, convolutional neural networks have been proposed for modelling temporal structure: Dieleman et al. [38] propose deep learning architectures for genre, artist and key classification tasks. Hamel et al. [24] propose a deep learning architecture incorporating multiple feature aggregation functions for tag classification.

The approach proposed in this work resembles methods applying statistical models to quantised feature sequences [30], [32]–[34]. In contrast, we propose to compute summary statistics directly from estimated sequential models. Since the obtained statistics may be compared using a metric, our approach has the potential to be combined with indexing and hashing schemes for computationally efficient retrieval [39]–[41], while retaining information on temporal structure. Our method of computing multiple representations using down-sampling resembles the approach proposed by Dieleman and

Schrauwen [36].

Note that our approach differs from Cilibrasi *et al.* [42], who propose pairwise sequence compressibility to quantify similarity. We did not pursue this approach for low-specificity tasks, based on results for the pairwise prediction approach reported in Section IV-A4. Note that we may take compression rates as estimates of sequential Shannon entropy rates, inviting further comparison or combination with related measures of sequential complexity [43]–[45]. Such measures have to date not been evaluated quantitatively in music content analysis, inviting further investigation beyond the scope of this work.

III. APPROACH

Assume that we are given the audio feature vector sequence $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_T)$. Similar to the descriptor proposed in [46], as a means of quantifying the sequential complexity of \mathbf{V} , we compute the compression rate $R_\lambda(\mathbf{V})$,

$$R_\lambda(\mathbf{V}) = \frac{C(\mathbf{V}, \lambda)}{T} \quad (2)$$

where $C(\mathbf{V}, \lambda)$ denotes the number of bits required to represent \mathbf{V} , given a quantisation scheme with λ levels and using a specified sequential compression scheme. To obtain a length-invariant measure of sequential complexity, we normalise with respect to the sequence length T .

Given the i th track in our collection, we compute compression rates for feature sequences extracted from musical audio. We refer to the set of compression rates as *feature complexity descriptors* (FCDs). For features based on constant frame rate, we compute FCDs using the original feature sequence, in addition to FCDs computed on downsampled versions of the original sequence; we consider downsampling factors 1, 2, 4, 8. We distinguish among temporal resolutions using the labels FCD1, FCD2, FCD4, FCD8, respectively. For features based on variable frame rate, we compute FCDs with no further downsampling applied.

Thus proposed, consider FCDs computed on a hypothetical scalar-valued feature sequence exhibiting a high amount of temporal structure, either due to periodicity or locally constant regions (Fig. 1 (a), (b)). For such sequences, we obtain low values for R_λ , since the quantised feature sequence may be encoded efficiently. Conversely, if we discard temporal structure by randomly shuffling the original feature sequence (Fig. 1 (c)), we obtain high values for R_λ , since the quantised feature sequence no longer admits an efficient encoding. In contrast to FCDs, feature moments such as mean and variance are invariant to any such re-ordering of features. We observe that feature moments have been widely applied for low-specificity content analysis tasks. Considering that FCDs have similar dimensionality to feature moments and assuming that temporal order of features is informative for our considered tasks, we therefore expect that FCDs may be used to improve prediction accuracy with respect to using feature moments alone, for our considered tasks.

A. Similarity rating prediction

For the task of similarity rating prediction, assume that we have a distance metric which we use to compare descriptor

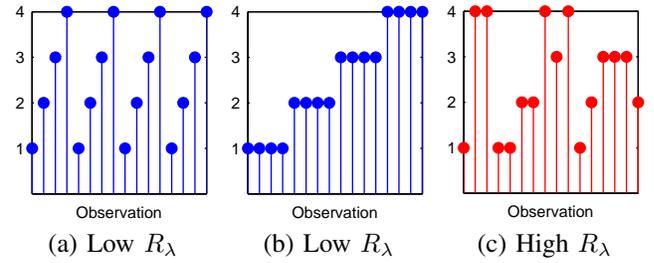


Fig. 1. Hypothetical sequences with low and high R_λ , assuming $\lambda = 4$.

vectors computed on pairs of tracks. We hypothesise that the pairwise distance between descriptors correlates with the similarity rating associated with track pairs. To predict similarity ratings we take as our feature space pairwise distances between descriptor vectors and apply multinomial regression. We use $\mathbf{r}_{i,n}$ to denote the n th descriptor vector computed for the i th track in our collection, with $1 \leq n \leq N$ and given a set of N available descriptor vectors. We compute separate descriptor vectors across audio features and across FCD resolutions, with each vector component in $\mathbf{r}_{i,n}$ corresponding to a quantisation granularity λ . We denote with $\mathbf{d}_{\langle i,j \rangle}$ the distances between $\mathbf{r}_{i,n}$, $\mathbf{r}_{j,n}$ obtained across all N descriptor vectors, using our assumed distance measure. Given the pair of tracks $\langle i, j \rangle$ whose similarity rating we seek to predict, we estimate the probability of similarity score $k \in [1..K]$ as

$$P(S = k | \mathbf{d}_{\langle i,j \rangle}) = \frac{\exp(\beta_k^T \mathbf{d}_{\langle i,j \rangle} + \gamma_k)}{\sum_{m=1}^K \exp(\beta_m^T \mathbf{d}_{\langle i,j \rangle} + \gamma_m)} \quad (3)$$

where β_k , γ_k are the model parameters associated with outcome k , given a total of K similarity scores. We predict similarity ratings by determining the value of k which maximises $P(S = k | \mathbf{d}_{\langle i,j \rangle})$. We describe our model estimation method in Section IV-A3.

B. Song year prediction

For the task of song year prediction, we hypothesise that descriptor values correlate with the chart entry date of tracks. Following [10] we apply a linear regression model. Given the i th track in our collection, we predict the associated chart entry date y_i using a linear combination of components in descriptor vectors $\mathbf{r}_{i,n}$,

$$\hat{y}_i = \sum_{n=1}^N \theta_n^T \mathbf{r}_{i,n} + \alpha \quad (4)$$

where θ_n denotes regression coefficients for the n th descriptor vector as specified for similarity rating prediction, and where α denotes the model intercept. We describe our model estimation method for song year prediction in Section IV-B1. We motivate use of both multinomial and linear regression techniques as a straightforward means of evaluating the utility of FCDs for determining similarity based on a metric space. We perform our evaluation by considering predictive accuracy, in addition to interpreting estimated coefficients as feature utilities.

IV. EVALUATION

For our evaluations, we use a collection of 15 473 entries from the American *Billboard Hot 100* singles popularity chart¹. Each entry in the dataset is represented by a track excerpt of approximately 30s of audio, and is annotated with a chart entry date. Chart entry dates span the years 1957–2010 ($M = 1982.9y$, $SD = 15.4y$).

For each track excerpt in the dataset, we extract a set of 25 audio features, using MIRToolbox [47] version 1.3.2 and using the framewise chromagram representation proposed by Ellis and Poliner [48]. With the exception of rhythmic features, which are computed using predicted onsets, features are based on a constant frame rate of 40Hz. Table I summarises the set of evaluated audio features.

Feature name	Description
Chroma (Ellis and Poliner)	12-component chromagram based on using phase-derivatives to identify tonal components in spectrum [48].
dynamics.rms	Root mean square of amplitude.
rhythm.tempo	Tempo estimate based on selecting peaks from autocorrelated onsets.
rhythm.attack.time	Duration of onset attack phase.
rhythm.attack.slope	Slope of onset attack phase.
spectral.centroid	First moment of magnitude spectrum.
spectral.brightness	Proportion of spectral energy above 1500Hz.
spectral.spread	Second moment of magnitude spectrum.
spectral.skewness	Skewness coefficient of magnitude spectrum.
spectral.kurtosis	Excess kurtosis of magnitude spectrum.
spectral.rolloff95	95th percentile of energy contained in magnitude spectrum.
spectral.rolloff85	85th percentile of energy contained in magnitude spectrum.
spectral.spectentropy	Shannon entropy of magnitude spectrum.
spectral.flatness	Wiener entropy of magnitude spectrum.
spectral.roughness	Average roughness [49] between peak pairs in magnitude spectrum.
spectral.irregularity	Squared amplitude difference between successive partials [50].
spectral.mfcc	12-component MFCCs [51] (excluding energy coefficient).
spectral.dmfcc	First-order differentiated MFCCs.
spectral.ddmfcc	Second-order differentiated MFCCs.
timbre.zerocross	Zero crossing rate.
timbre.spectralflux	Half-wave rectified L1 distance between magnitude spectrum at successive frames [52].
tonal.chromagram.centroid	Centroid of 12-component chromagram.
tonal.keyclarity	Peak correlation of chromagram with key profiles [53].
tonal.mode	Predicted mode after correlating chromagram with key profiles.
tonal.hcdf	Flux of 6-dimensional tonal centroid [54].

TABLE I
SUMMARY OF EVALUATED AUDIO FEATURES.

In addition to FCDs, for each track excerpt we compute the mean and standard deviation, based on frame-level representation with no downsampling applied. We refer to the latter non-sequential descriptors as feature moment descriptors (FMDs). We compute FCDs as described in Section III, where for the case of the vector-valued features chroma, MFCCs and delta-MFCCs we apply principal component analysis (PCA)

in track-wise fashion as a preliminary decorrelation step. We then quantise and compress each resulting component separately, before averaging obtained compression lengths across components. We apply PCA, since we seek to quantify temporal structure in feature vector sequences while disregarding any correlation among feature vector components. We quantise features by applying equal-frequency binning with $\lambda \in \{3, 4, 5\}$ levels; we perform relatively coarse quantisation to ensure that each symbol occurs frequently, regardless of downsampling factor.

We choose equal-frequency binning to ensure that obtained strings have a consistent stationary distribution; the obtained compression rates therefore are a function of temporal structure alone. The value $\log \lambda$ may be interpreted as the theoretical compression rate for a temporally uncorrelated sequence. We compress symbol sequences using the *prediction by partial match* (PPM) algorithm², described in [55]. We consider PPM a general-purpose string compression algorithm which may be substituted with an alternative compressor; in initial experiments we obtained similar results using Lempel-Ziv compression [31]. Nevertheless, we note that PPM compresses efficiently compared to alternative compression schemes [55]. We set the PPM model order to 5 symbols, based on the observation that for uncorrelated sequences, distinct substrings of length 5 are unlikely to occur frequently.

With a view to characterising the feature space represented by FCDs, we perform a track-wise exploratory analysis of computed FCDs. For each track excerpt in our collection, we compute FCDs based on MFCC features alone. We obtain a scalar-valued score for each excerpt by averaging FCDs across quantisation levels λ and across temporal resolutions. Next, across artists in our collection we compute the median of obtained FCD scores. To facilitate interpretation, we consider only artists with a minimum number of 20 chart entries; thus out of 5 455 artists in our collection we consider 129 artists. We then rank artists according to median FCD scores. Shown in Table II, we report the 20 lowest-ranking and highest-ranking artists. Additionally, across artists we report as medoid tracks those tracks whose FCD score minimises the error with respect to the median.

Comparing track groups, the lowest-ranking artists are predominantly vocalists with repertoire of jazz ballads and slow-moving pieces (e.g. Johnny Mathis, Barbara Streisand). In contrast, the artists with highest complexity values stand for music with strong percussive and aggressive components, from up-tempo surf-rock (Jan & Dean), through 1980s Power Rock (Van Halen) and Hip Hop (Eminem). Informal listening to medoid tracks supports this observation, with the exception of the medoid track by artist Etta James. We view this observation in support of our expectation that FCDs may be useful for low-specificity similarity and subsequently demonstrate validity of our expectation for the similarity tasks considered in this work. Note however that we make no claim that FCDs capture any notion of musical complexity as proposed in [56]. While beyond the scope of this paper, track-wise analysis of FCDs merits further investigation.

¹<http://www.billboard.com>

²<http://www.cs.technion.ac.il/~ronbeg/vmm/index.html>

Lowest-ranking scores			Highest-ranking scores		
Score	Artist name	Medoid track name	Score	Artist name	Medoid track name
1.223	Johnny Mathis	Starbright	1.286	Jan & Dean	The Anaheim ... Association
1.234	Barbra Streisand	Didn't We	1.286	Bryan Adams	This Time
1.240	The Platters	Trees	1.287	Eric Clapton	After Midnight
1.245	Bobby Vinton	Rain Rain Go Away	1.287	Creedence Clearwater Revival	Who'll Stop The Rain
1.247	Connie Francis	(He's My) Dreamboat	1.287	The Rolling Stones	Tell Me (You're Coming Back)
1.251	Andy Williams	Sweet Memories	1.288	Johnny Cash	It's Just About Time
1.252	Jim Reeves	I Guess I'm Crazy	1.288	Chubby Checker	Whole Lotta Shakin' Goin' On
1.256	John Denver	Sweet Surrender	1.288	The Kinks	Tired Of Waiting For You
1.256	Barry Manilow	I Write The Songs	1.288	Eddie Money	Maybe I'm A Fool
1.261	Johnny Tillotson	I Rise, I Fall	1.288	Aerosmith	Hole In My Soul
1.261	Dionne Warwick	If We Only Have Love	1.288	Van Halen	When It's Love
1.261	Helen Reddy	Delta Dawn	1.289	The Doobie Brothers	What A Fool Believes
1.262	Etta James	Seven Day Fool	1.289	Marvin Gaye	Pretty Little Baby
1.263	Carpenters	Touch Me When We're Dancing	1.289	Madonna	Secret
1.263	Frank Sinatra	Talk To Me	1.290	Paul Revere & The Raiders	Country Wine
1.264	Engelbert Humperdinck	In Time	1.291	James Brown	Signed, Sealed, And Delivered
1.264	Brenda Lee	Too Many Rivers	1.291	Janet Jackson	Black Cat
1.264	Nat King Cole	Nothing In The World	1.291	The Isley Brothers	Harvest For The World
1.266	Gene Pitney	Town Without Pity	1.293	Freddy Cannon	Muskrat Ramble
1.267	Tom Jones	With These Hands	1.297	Eminem	Cleanin' Out My Closet

TABLE II

ARTISTS RANKED ACCORDING TO MEDIAN TRACK-WISE FCD SCORE. FOR EACH ARTIST, FCDs AVERAGED ACROSS QUANTISATION LEVELS λ AND ACROSS TEMPORAL RESOLUTIONS, USING MFCCs AS AUDIO FEATURE. TABLE REPORTS LOWEST-RANKING AND HIGHEST-RANKING SCORES.

A. Similarity rating prediction

We evaluate similarity rating prediction using annotations collected for a subset of the chart music dataset. Prior to our investigations, we obtained a total of 7784 pairwise similarity ratings from 456 subjects participating in a web-based listening test³. Subjects were asked to quantify pairwise musical similarity between successive pairs of track excerpts using a five-point ordinal scale, with score ‘1’ corresponding to ‘not similar’ and score ‘5’ corresponding to ‘very similar’. We assume that subjects have an internal similarity scale which they use to perform ratings. Therefore, we omit any training step from the rating process. Note that while we prescribe that pairwise similarity ratings are made using a five-point scale, we do not assume that similarities are judged using an absolute scale across listeners. Given three track pairs for which we have respective ratings (4, 5), (5, 5), (1, 2), we view the ratings as quantifying relative agreement, compared to (4, 1), (5, 1), (1, 4).

For human similarity judgements, two issues prompt consideration: In addition to music being inherently subjective [57], human similarity judgements are context-dependent [58], [59]. We motivate our assumption of an internal similarity scale on the basis that Western popular music is widely disseminated and that listeners might form similarity judgements using a common factor. We verify our assumptions by quantifying similarity rating agreement.

When presenting track pairs to listeners, we select the first song in each pair using uniform sampling. For the second song in each pair, we again apply uniform sampling, however we bias towards proximate chart entry times by restricting the permissible chart entry deviation to $\leq 1y$ with probability 0.9. We bias as a means of controlling for historical changes in audio production, which might affect similarity ratings [60]. We obtain a median of 6 ratings per subject, with each

rating corresponding to a unique track pair. Table III displays obtained score counts.

As shown in Table III, the majority of ratings are associated with scores less than ‘3’, corresponding to relative dissimilarity on the five-point scale. We contend that for music content analysis based on an ensemble of systems as proposed in [61], the entire target set of predicted musical similarity might be used when forming recommendations. In contrast, for track recommendation relying on predicted similarity alone, when forming recommendations, it is typically of interest to consider tracks deemed similar to a query, while disregarding tracks deemed dissimilar [62]. Pertaining to the first use case, we perform evaluations using the five-point scale ratings, as defined previously. Pertaining to the second use case, we merge similarity ratings with scores ‘1’ and ‘2’, thus discarding any distinction between similarity ratings with low scores. We then perform our evaluations using the resulting four-point scale ratings.

Count	Similarity score				
	1	2	3	4	5
	2060	2115	1742	1391	476

TABLE III

SIMILARITY SCORE COUNTS OBTAINED FROM WEB-BASED LISTENING TEST.

To assess the consistency of similarity ratings, we collected an additional set of similarity ratings under controlled experimental conditions, involving 12 subjects aged 21y–42y. Subjects were assessed using the Ollen musical sophistication index (OMSI) [63]. We obtain a median OMSI score of 241, with an associated median of 0.75 years of formal musical training. To avoid subject fatigue, we imposed no minimum number of ratings per subject, and collected ratings during two 30-minute sessions. We selected stimuli by sampling uniformly from the set of track pairs for which we have

³<http://webprojects.eecs.qmul.ac.uk/matthiasm/audioquality-pre/check.php>

prior ratings. Across subjects, we obtain a median of 42 ratings ($M = 45.4$, $SD = 29.3$). We aggregate controlled-condition ratings across subjects and thus obtain a total of 509 controlled-condition similarity ratings, corresponding to 6.5% coverage of web-based similarity ratings. Table IV displays a confusion matrix of web-sourced versus controlled-condition similarity ratings.

		Controlled-condition				
		1	2	3	4	5
Web-sourced	1	64	34	17	10	0
	2	55	44	18	14	4
	3	26	41	26	25	5
	4	16	30	16	24	7
	5	6	9	5	8	5

TABLE IV
CONFUSION MATRIX OF WEB-SOURCED VERSUS
CONTROLLED-CONDITION SIMILARITY RATINGS.

We quantify the agreement between controlled-condition and web-sourced similarity ratings. We report results for both five-point and four-point rating scales; for each agreement statistic we report results for the four-point rating scale in brackets. We first quantify agreement using Kendall's correlation coefficient τ_b , as defined in (5). We obtain a correlation of 0.274 (0.250), with $p < 0.001$ based on a permutation test for the hypothesis of no correlation. We then compute a confidence interval for the obtained sample correlation by applying bootstrap sampling [64]. At the 95% level, we obtain correlations in the range [0.205, 0.337] ([0.173, 0.325]). Subsequently, we consider the correlation 0.337 (0.325) an upper bound on attainable accuracy using our proposed method of similarity rating prediction. As a second measure of rating agreement, we compute Spearman's correlation coefficient ρ_s , where we obtain 0.329 (0.278) for ratings aggregated across subjects. Analogously by applying bootstrap sampling, at the 95% level we obtain correlations in the range [0.247, 0.404] ([0.193, 0.361]). We consider the correlation 0.404 (0.361) an upper bound on attainable accuracy based on ρ_s . Finally, using Table IV and interpreting the controlled-condition rating process as a multinomial classification task, we obtain a balanced classification accuracy (BA) of 0.292 (0.345); the corresponding 95% confidence interval is [0.254, 0.336] ([0.304, 0.393]).

1) *Distance measures*: We predict similarity ratings by applying multinomial regression to pairwise Euclidean distances between descriptor vectors, using the approach described in Section III-A. As an additional baseline distance measure, using (1) and assuming Gaussianity and diagonal covariance, we compute the KLD on pairs of FMDs. We logarithmically transform distances obtained using the KLD, which we observed improved prediction accuracy.

As a baseline distance accounting for temporal structure, we compute the cross-prediction error between audio feature sequences, with each feature sequence represented at the original frame level. Following [65], we apply state space embedding [66] separately to pairs of feature sequences. Given feature vectors ($\mathbf{v}_1, \dots, \mathbf{v}_T$) each with dimensionality h , state space embedding produces higher-dimensional feature

vectors with dimensionality dh by stacking d consecutive vectors $\mathbf{v}_{t-d}, \dots, \mathbf{v}_{t-1}$ at each time step t . We perform cross-predictions by determining sequential successors of nearest neighbours in the embedded space, using the approach given in [67]. As a distance measure between predicted and observed feature sequences, we compute the normalised mean square error [65]. We consider parameter $d \in \{8, 12, 16, 20\}$ and report results for $d = 12$, which yields highest average correlation between computed pairwise distances and similarity annotations. We apply square-root transformation to pairwise distances, which we observed improved similarity rating prediction accuracy.

2) *Performance statistics*: To quantify the accuracy of similarity rating prediction, as discussed in [68] we compute Kendall's τ_b and Spearman's ρ_s , both which are ordinal measures of association between predicted and annotated similarity ratings. We define Kendall's τ_b as follows. Assume that we have sequences $\mathcal{Q} = (q_1, \dots, q_M)$, $\mathcal{O} = (o_1, \dots, o_M)$. The pair $d_{i,j} = ((q_i, o_i), (q_j, o_j))$ is termed *concordant*, if $q_i > q_j$ and $o_i > o_j$, or if $q_i < q_j$ and $o_i < o_j$. Analogously, $d_{i,j}$ is termed *discordant*, if $q_i < q_j$ and $o_i > o_j$, or if $q_i > q_j$ and $o_i < o_j$. Kendall's τ_b is defined as

$$\tau_b = \frac{M_c - M_d}{\sqrt{(M_p - M_q)(M_p - M_o)}} \quad (5)$$

where M_c , M_d respectively denote the number of concordant and discordant pairs and where $M_p = \frac{1}{2}M(M-1)$ denotes the total number of pairs. Terms M_q , M_o respectively denote the number of pairs with tied (q_i, q_j) and with tied (o_i, o_j) . In the denominator, the normalisation is with respect to the geometric mean of adjusted pair counts $(M_p - M_q)$, $(M_p - M_o)$. Yielding values in the range $[-1, 1]$, τ_b may be interpreted as an estimate of the difference in probability of sampling a concordant pair versus sampling a discordant pair in $(\mathcal{Q}, \mathcal{O})$, while accounting for ties.

As a second measure of prediction accuracy, we compute Spearman's ρ_s , corresponding to the product-moment correlation coefficient between separately ranked \mathcal{Q} , \mathcal{O} . We assign unique ranks to tied values, before computing average ranks across tied values. Note that in contrast to τ_b , the value of ρ_s is a function of assigned ranks. Thus, in the presence of ties τ_b may be viewed as a more appropriate means of comparing ordinal sequences [69]. Nevertheless, we compute ρ_s , since its square yields a direct interpretation as proportion of explained variance between assigned ranks.

As a third performance measure, we view our prediction task as multinomial classification and compute BA. Note that in contrast to τ_b , ρ_s , BA disregards the ordering of rating scores. Based on the notion of rating agreement given in Section IV-A, we thus consider BA a subsidiary measure of performance compared to τ_b , ρ_s .

3) *Model estimation*: We evaluate similarity rating prediction by applying hold-out validation to web-sourced annotations. We use 60% of annotations for training, with the remainder of annotations used for testing.

We apply multinomial regression separately to sets of distances between descriptor vectors, as specified in Table V. We standardise distances by subtracting the mean and dividing

by the variance of the training data. Note that we compute FCD vectors separately across temporal resolutions and across audio features. Based on a set of 25 audio features, given a pair of tracks we thus obtain a total of 100 distances between compression-based descriptor vectors. Furthermore, note that when combining sets of descriptors we aggregate among obtained distances. Thus given a pair of tracks, when combining sets 1, 3, 4 as specified in Table V, we obtain 150 distances. As given in (3), we weight distances individually.

In our training step, we estimate multinomial regression parameters using elastic net regularisation (ENR) [70] based on coordinate descent⁴ [71]. We denote with $\beta = (\beta_1^T, \dots, \beta_K^T)^T$, $\gamma = (\gamma_1, \dots, \gamma_K)^T$ regression coefficients and model intercepts as given in (3). Using ENR, we solve

$$\min_{\beta, \gamma} \left\{ \eta \left(\nu \|\beta\|_1 + (1 - \nu) \frac{1}{2} \|\beta\|_2^2 \right) - \ell(\beta, \gamma) \right\} \quad (6)$$

where $\ell(\beta, \gamma)$ denotes model log-likelihood. Furthermore, η and ν respectively are *shrinkage* and *elastic net penalty* parameters, with $\eta > 0$ and $0 \leq \nu \leq 1$. Thus, ν determines the relative contribution of regularisation due to L1 and L2 norms, whereas η scales the regularisation penalty. For each performance statistic given in Section IV-A2 and for each rating scale as given in IV-A, we apply hold-out validation to training data and optimise η by determining maximal prediction accuracy. We consider ν a hyper-parameter which we assign constant value; we optimise Kendall's τ_b with respect to the five-point rating scale and using a model incorporating FCDs and FMDs, where we again apply hold-out validation to training data.

Chroma (Ellis and Poliner)	0.04*	0.04*	0.03*	0.17*	0.18*	0.15*	0.08*
dynamics.rms	0.09*	0.08*	0.09*	0.11*	0.10*	0.09*	0.07*
rhythm.temp	0.07*	0.04*	0.02	0.01	0.01	0.01	0.01
rhythm.attack.time	0.05*	0.05*	0.03*	0.03	0.03	0.03	0.03
rhythm.attack.slope	0.03	0.03*	0.01	0.05*	0.05*	0.05*	0.05*
spectral.centroid	0.12*	0.10*	0.05*	0.07*	0.06*	0.06*	0.03*
spectral.brightness	0.12*	0.13*	0.09*	0.09*	0.08*	0.05*	0.02
spectral.spread	0.11*	0.10*	0.08*	0.07*	0.10*	0.07*	0.04*
spectral.skewness	0.04*	0.06*	0.07*	0.11*	0.10*	0.08*	0.03*
spectral.kurtosis	0.03	0.06*	0.06*	0.10*	0.10*	0.07*	0.05*
spectral.rolloff95	0.08*	0.06*	0.04*	0.08*	0.07*	0.05*	0.05*
spectral.rolloff85	0.11*	0.09*	0.05*	0.08*	0.07*	0.04*	0.03
spectral.spectentropy	0.13*	0.12*	0.09*	0.07*	0.08*	0.06*	0.03*
spectral.flatness	0.09*	0.08*	0.04*	0.07*	0.06*	0.05*	0.04*
spectral.roughness	0.03	0.04*	0.06*	0.09*	0.10*	0.07*	0.04*
spectral.irregularity	0.05*	0.06*	0.07*	0.10*	0.11*	0.07*	0.03*
spectral.mfcc	0.14*	0.15*	0.07*	0.18*	0.19*	0.15*	0.08*
spectral.dmfcc	0.14*	0.16*	0.03	0.08*	0.04*	0.06*	0.05*
spectral.ddmfcc	0.14*	0.16*	0.03	0.04*	0.01	0.04*	0.04*
timbre.zerocross	0.12*	0.11*	0.07*	0.05*	0.06*	0.04*	0.02
timbre.spectralflux	0.19*	0.18*	0.04*	0.09*	0.08*	0.04*	0.04*
tonal.chromagram.centroid	0.10*	0.10*	0.12*	0.07*	0.07*	0.07*	0.03*
tonal.keyclarity	0.15*	0.15*	0.13*	0.14*	0.11*	0.07*	0.01
tonal.mode	0.08*	0.10*	0.09*	0.15*	0.13*	0.07*	0.01
tonal.hcdf	0.11*	0.12*	0.03*	0.09*	0.05*	0.03	0.02
FMDs (Euclidean)				FCD1	FCD2	FCD4	FCD8
FMDs (KLD)							
Frame-level cross-prediction							

Fig. 2. Feature-wise absolute correlation $|\tau_b|$ between pairwise distances and web-sourced similarity annotations. Pairwise distances respectively obtained using FMDs compared using Euclidean distance and KLD (first and second columns), cross-prediction (third column), Euclidean distance applied to FCDs (remaining columns). Starred entries indicate significance, where we apply Bonferroni correction to $\alpha = 0.05$.

4) *Results*: We examine the correlation between descriptor distances and five-point scale similarity ratings across individual audio features. Fig. 2 depicts correlations τ_b for FCDs and FMDs, where we compare FMDs using both Euclidean distance and KLD. In addition to FMDs, as described in Section IV-A1 we consider as a baseline the cross-prediction error.

We observe that FCDs and FMDs both yield maximum correlation 0.19 (comparing FCD2 to FMDs, with both distances computed using Euclidean distance); similarly, FMDs compared using KLD yield maximum correlation 0.18. Across descriptors, with $\alpha = 0.05$ and applying Bonferroni correction, the majority of features yield significant correlations. In contrast, for cross-prediction, effect sizes are comparatively small. Comparing descriptors, for FCD2 we observe correlations exceeding 0.1 for 9 features, and for 12 features for the case of FMDs compared either using KLD or Euclidean distance. On average, FMDs yield greater correlation compared to FCD1 (0.095 versus 0.087). However, for specific features FCDs yield higher correlation than FMDs. Comparing FCDs amongst temporal resolutions, we observe a monotonically decreasing relationship between downsampling factor and average correlation.

Fig. 3 displays a comparison of similarity rating prediction accuracy, where for each descriptor set in Table VI we apply feature selection as described in Section IV-A3. We estimate models using τ_b , ρ_s , BA as performance statistics. We consider both 5-point and 4-point rating scales. In particular, we consider the performance gain obtained by including FCDs in our models.

Across both rating scales, we observe that FCDs are outperformed by FMDs compared using KLD alone, or using Euclidean distance and KLD in combination. However, a combination of FCDs and FMDs outperforms evaluated combinations employing FMDs alone. By incorporating compression descriptors, compared to FMDs based on aggregated KLD and Euclidean distance, based on the five-point rating scale we obtain absolute performance gains of 0.033, 0.030, 0.013 with respect to ρ_s , τ_b , BA. The respective relative performance gains are 10.4%, 11.3%, 4.7%. Based on the four-point rating scale, we obtain absolute performance gains of 0.059, 0.051, 0.021; the respective relative performance gains are 31.1%, 29.1%, 7.2%. For the model using ρ_s and the four-point rating scale, Table VII displays confusion matrices of predicted versus annotated ratings. We test for differences between correlations by applying bootstrap sampling to predicted and observed similarity ratings, from which in turn we estimate standard errors of performance statistics. Based on a one-way analysis of variance with Tukey-Kramer post-hoc analysis and setting $\alpha = 0.05$, we reject the hypothesis of no difference between correlations across all considered pairs, for all considered performance statistics.

Fig. 4 displays regression coefficients across features and descriptor classes, where we consider the best-performing model evaluated in Fig. 3 based on ρ_s and using the five-point rating scale. We sum regression coefficient magnitudes across each of the K binary classifiers given in (3), before normalising the obtained values to sum to one. Comparing

⁴http://www.stanford.edu/~hastie/glmnet_matlab/

Set	Track representation	Descriptor vector components	Distance measure	Prediction coeffs.
1.	FCDs	$\lambda \in \{3, 4, 5\}$	Euclidean	4×25
2.	Frame sequence	N/A	Cross-prediction error	25
3.	FMDs	Mean, Std	Euclidean	25
4.	FMDs	Mean, Var	KLD	25
5.	Combine 3, 4			50
6.	Combine 1, 3, 4			150

TABLE V

SUMMARY OF DESCRIPTOR COMBINATIONS EVALUATED FOR SIMILARITY RATING PREDICTION. THIRD COLUMN DENOTES COMPONENTS INCLUDED IN DESCRIPTOR VECTORS. FIFTH COLUMN LISTS NUMBER OF COEFFICIENTS IN MULTINOMIAL REGRESSION MODEL (EXCLUDING INTERCEPTS).

Set	Track representation	Descriptor vector components	Prediction coeffs.
1.	FMDs	Mean, Std	$21 \times 2 + 4 \times 24$
2.	FCDs	$\lambda \in \{3, 4, 5\}$	$25 \times 4 \times 3$
3.	Combine 1, 2		

TABLE VI

SUMMARY OF DESCRIPTOR COMBINATIONS EVALUATED FOR SONG YEAR PREDICTION. FOURTH COLUMN LISTS NUMBER OF COEFFICIENTS IN LINEAR REGRESSION MODEL (EXCLUDING INTERCEPT).

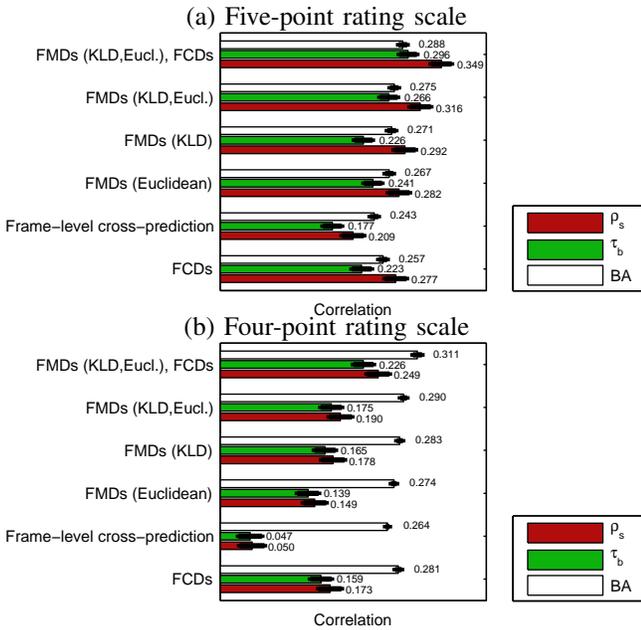


Fig. 3. Similarity rating prediction accuracy. Standard errors obtained by bootstrap sampling pairs of predicted and observed similarity ratings.

Annotated	(a) FMDs (KLD, Eucl.)				(b) FMDs (KLD, Eucl.), FCDs			
	Predicted 1;2	3	4	5	Predicted 1;2	3	4	5
1;2	1490	96	80	1	1361	152	131	23
3	532	75	78	0	458	115	101	11
4	435	48	87	4	311	111	129	23
5	130	15	42	0	106	37	37	7

TABLE VII

CONFUSION MATRICES OF PREDICTED VERSUS ANNOTATED SIMILARITY RATINGS, FOR MODEL BASED ON FOUR-POINT RATING SCALE AND ρ_s .

FMDs and FCDs, we observe that both FCDs and FMDs are selected within individual features. FCDs appear to be selected across diverse temporal resolutions, with emphasis on higher temporal resolutions. We observe that multiple FCD resolutions are selected within the same feature.

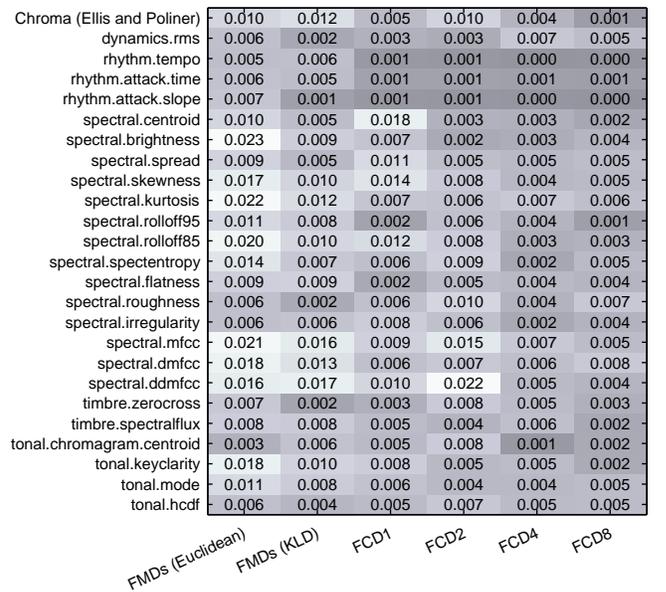


Fig. 4. Normalised regression coefficient magnitudes, estimated using elastic net regression, for task of similarity rating prediction. Candidate descriptor set comprised of FCDs compared using Euclidean distance, and FMDs compared using Euclidean distance and KLD.

B. Song year prediction

For song year prediction, we compute FCDs and FMDs as performed for similarity rating prediction. We use chart entry dates as our annotation data and apply the linear regression model given in (4). Fig. 5 displays a histogram of chart entry dates.

1) *Model estimation:* To evaluate our descriptors for song year prediction, we partition the dataset into random training and testing subsets, where we ensure that title or artist strings are not duplicated across subsets. We apply the aforementioned filtering procedure to control for potential cover version and album effects, in addition to any analogous effects at the level of artists [72]. The resulting training and testing datasets consist of 10728 and 4745 tracks respectively. We deem as outliers descriptor values in the training data exceeding 10 standard

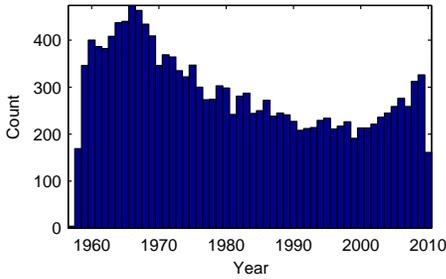


Fig. 5. Histogram of chart entry dates.

deviations beyond the 99th percentile. We replace such outliers with imputed values, using the K -nearest neighbour algorithm.

We apply linear regression separately to sets of descriptor vectors, as specified in Table VI. We standardise descriptors by subtracting the mean and dividing by the variance of the training data. As performed for similarity rating prediction, we compute FCDs separately across temporal resolutions and across audio features. In contrast, we apply linear regression directly to descriptor vectors without the intermediate step of computing distances. Based on a set of 25 audio features, given a single track we obtain a total of 300 scalar-valued FCDs, for each of which we estimate a single regression coefficient. Note that since we represent FMDs using the mean and standard deviation, we estimate two regression coefficients for each univariate audio feature. For FMDs, it follows that we estimate 24 regression coefficients for MFCCs and chroma features.

As was performed for similarity rating prediction, we estimate linear regression parameters using ENR. We denote with $\theta = (\theta_1^T, \dots, \theta_N^T)^T$, α regression coefficients and the model intercept as given in (4). Using ENR, we solve

$$\min_{\theta, \alpha} \left\{ \eta \left(\nu \|\theta\|_1 + (1 - \nu) \frac{1}{2} \|\theta\|_2^2 \right) + \text{SSR}(\theta, \alpha) \right\} \quad (7)$$

where $\text{SSR}(\theta, \alpha)$ denotes the sum of squared residuals. Both η , ν behave as defined in (6). We apply cross-validation to training data and optimise η by determining minimal prediction mean square error. We again consider ν a hyper-parameter which we assign constant value; we optimise prediction mean square error based on a model incorporating FCDs and FMDs, and by applying cross-validation to training data. We threshold predictions to fall in the range [1957y .. 2010y].

In addition to the year prediction task based on individual tracks, we evaluate prediction performance when considering groups of tracks. We perform this experiment to establish whether FCDs consistently improve performance when combined with grouped FMDs, or if grouped FMDs amortise any potential performance gain due to FCDs. We select groups of tracks by applying a non-overlapping sliding window to chart entry dates. We then take as descriptor vector $\mathbf{r}'_{w,n}$ the average

$$\mathbf{r}'_{w,n} = \frac{1}{|C_w|} \sum_{i \in C_w} \mathbf{r}_{i,n} \quad (8)$$

where C_w denotes the set of tracks at window position w . We apply the windowing procedure separately to training and testing data sets. Note that by windowing tracks, at each window position we assume prior knowledge of differences among

chart entry times in training and testing data, respectively. For a given window size, we average descriptor vectors in the training data and proceed as described in Section IV-B1. Given the obtained regression model and given averaged descriptor vectors at window position z in the testing data, we seek to predict the associated window centre y'_z .

2) *Performance statistics*: We quantify prediction accuracy with respect to annotated chart entry dates, using the mean absolute error (MAE) and root mean square error (RMSE) statistics.

3) *Results*: Fig. 6 displays the result of exploratory analysis for song year prediction, where for FMDs and FCDs we group descriptor values across time, by applying a non-overlapping 2-year sliding window to chart entry dates. We restrict analysis to obtained spectral spread features [47]. The resulting year-wise box plots suggest that the examined descriptors are non-stationary with respect to chart entry dates, exhibiting distinct trends. To examine the behaviour of descriptors at a finer time scale, we apply a non-overlapping 30-day sliding window to chart entry dates, where at each window position we compute the mean descriptor value. Examining the sample autocorrelation of the resulting time series for lags in the range [1..15], we observe weaker correlations for FCDs compared to FMDs. Yet, both autocorrelations exhibit slowly decaying autocorrelations (Fig. 7), characteristic of a non-stationary time series [73]. Following the method of Box and Jenkins [74], we attempt to attain stationarity by applying first-order differencing to the time series. However, we observe autocorrelation close to -0.5 at unit lag, suggesting that the time series have been overdifferenced [73]. We interpret these observations as evidence for a non-trivial, trend-exhibiting process governing observed descriptor values [75].

Set	MAE	RMSE
FCDs	9.44 ± 0.096	11.54 ± 0.107
FMDs	8.28 ± 0.092	10.45 ± 0.113
Combined	7.38 ± 0.085	9.43 ± 0.107

TABLE VIII
SUMMARY OF SONG YEAR PREDICTION ACCURACY, EXPRESSED USING MAE AND RMSE STATISTICS. STANDARD ERRORS OBTAINED BY BOOTSTRAP SAMPLING PAIRS OF PREDICTED AND OBSERVED CHART ENTRY DATES.

Table VIII summarises the accuracy of song year prediction using MAE and RMSE statistics. Quantified using either MAE or RMSE, song year prediction based on FMDs outperforms prediction using FCDs alone. However, we observe that a combination of FMDs and FCDs yields the highest prediction accuracy. By incorporating FCDs we observe performance gains of 10.9%, 9.8% relative to FMDs, in terms of MAE and RMSE. As performed in Section IV-A4, we test for differences among prediction accuracies by applying bootstrap sampling to predicted and observed chart entry times, from which we estimate standard errors of MAE and RMSE statistics. Again using one-way analysis of variance with Tukey-Kramer post-hoc analysis and setting $\alpha = 0.05$, we reject the hypothesis of no difference between prediction accuracies across all pairs, for both MAE and RMSE.

Fig. 8 displays regression coefficients obtained using windowed chart entry dates. We compute coefficient magnitudes and normalise to sum to one. Thus computed, we interpret coefficient magnitudes as predictive utilities across individual audio features. In addition, we consider the utility of FCDs across time scales, compared to FMDs. Summed across features, we observe that compared to FCD1, FMDs are weighted more strongly (0.591 versus 0.201). Further examining relative weightings, we observe a prevalence of weight assigned to FCD1 compared to higher downsampling factors. However, we observe that individual features may be weighted relatively strongly across multiple temporal scales. Note from Table V that for chroma features, MFCCs and derivatives, FMD weights are summed across 24 prediction coefficients, compared to 3 coefficients for FCDs.

In Fig. 9 we examine prediction accuracy in response to windowed descriptors, as described in Section IV-A3 and quantified using MAE. For increasing window size up to 60d, performance improves monotonically across all considered descriptor sets. Across considered window sizes, using combined FCDs and FMDs we observe a mean performance gain of 17.5%, relative to using FMDs alone.

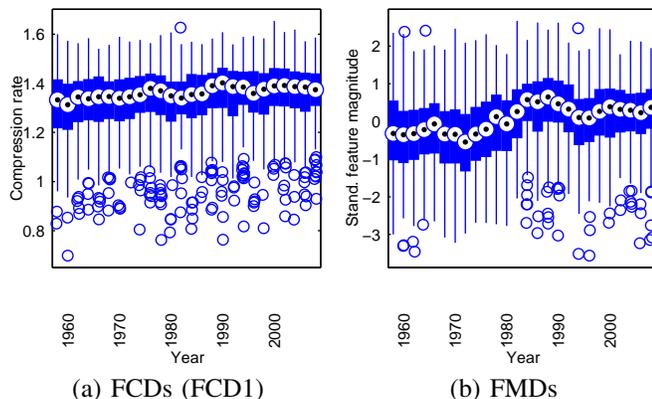


Fig. 6. Box plots of FCDs and FMDs computed using spectral spread features, with FCDs computed without downsampling. Each box corresponds to the position of a non-overlapping 1-year window applied to chart entry dates.

V. CONCLUSIONS

We have considered the problem of determining musical similarity, using feature sequences extracted from musical audio. In particular, we have considered musical similarity in the context of two low-specificity content retrieval tasks, namely similarity rating prediction and song year prediction. To this end, we have evaluated the utility of sequential complexity as a descriptor for quantifying musical similarity.

For both considered tasks, we observe that sequential complexity descriptors predict the outcome variable. Furthermore, in combination with feature moment descriptors, sequential complexity descriptors improve prediction accuracy with respect to the baseline. The results confirm that our proposed descriptors capture musically relevant information and that temporal structure is relevant in our chosen domain. Consequently, our results show that sequential complexity may

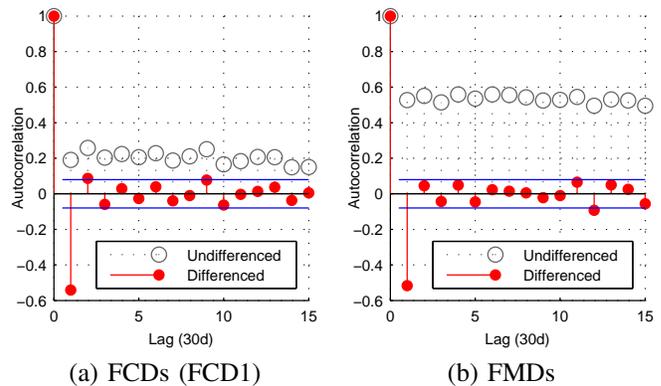


Fig. 7. Sample autocorrelation of undifferenced and differenced FCD, FMD averages. Descriptor averages obtained by applying non-overlapping 30-day window to chart entry dates. Descriptors computed on spectral spread features, with FCDs computed without downsampling. Horizontal bars indicate 95% confidence intervals under the assumption of Gaussian white noise for differenced time series.

Chroma (Ellis and Poliner)	0.0167	0.0023	0.0085	0.0036	0.1436
dynamics.rms	0.0135	0.0080	0.0022	0.0007	0.0043
rhythm.tempo	0.0013	0.0013	0.0012	0.0013	0.0012
rhythm.attack.time	0.0006	0.0008	0.0007	0.0009	0.0089
rhythm.attack.slope	0.0014	0.0014	0.0014	0.0015	0.0086
spectral.centroid	0.0039	0.0039	0.0051	0.0017	0.0245
spectral.brightness	0.0042	0.0004	0.0009	0.0018	0.0143
spectral.spread	0.0266	0.0061	0.0028	0.0019	0.0062
spectral.skewness	0.0076	0.0017	0.0017	0.0002	0.0110
spectral.kurtosis	0.0121	0.0034	0.0044	0.0038	0.0141
spectral.rolloff95	0.0010	0.0088	0.0016	0.0015	0.0334
spectral.rolloff85	0.0134	0.0044	0.0008	0.0015	0.0324
spectral.spectentropy	0.0031	0.0014	0.0061	0.0011	0.0104
spectral.flatness	0.0029	0.0034	0.0009	0.0015	0.0060
spectral.roughness	0.0237	0.0156	0.0099	0.0026	0.0064
spectral.irregularity	0.0070	0.0030	0.0001	0.0025	0.0159
spectral.mfcc	0.0176	0.0030	0.0029	0.0040	0.0741
spectral.dmfcc	0.0096	0.0040	0.0021	0.0016	0.0593
spectral.ddmfcc	0.0060	0.0018	0.0041	0.0023	0.0194
timbre.zerocross	0.0014	0.0022	0.0002	0.0012	0.0446
timbre.spectralflux	0.0114	0.0035	0.0042	0.0033	0.0034
tonal.chromagram.centroid	0.0022	0.0043	0.0003	0.0015	0.0178
tonal.keyclarity	0.0062	0.0060	0.0021	0.0003	0.0114
tonal.mode	0.0039	0.0028	0.0008	0.0020	0.0045
tonal.hcdf	0.0037	0.0009	0.0017	0.0019	0.0156
	FCD1	FCD2	FCD4	FCD8	FMD

Fig. 8. Normalised regression coefficient magnitudes, estimated using elastic net regularisation, for task of song year prediction. Candidate descriptor set comprised of FCDs and FMDs.

be used to improve the accuracy of low-specificity content retrieval based on bag-of features approaches.

Our proposed descriptors are computed in an unsupervised manner and may be implemented efficiently, requiring $O(n)$ time complexity for each track [76]. In addition, our proposed descriptors have similar dimensionality compared to feature moment descriptors. Since our descriptors may be computed off-line or incrementally and thereafter combined with indexing methods as proposed in [39]–[41], we deem them potentially applicable in large-scale content retrieval systems.

Similar to results obtained in [24], [35], [36], [77], our results using sequential complexity descriptors suggest that an approach based on multiple temporal resolutions is advantageous for determining musical similarity. As an alternative to downsampled features, we initially employed beat-

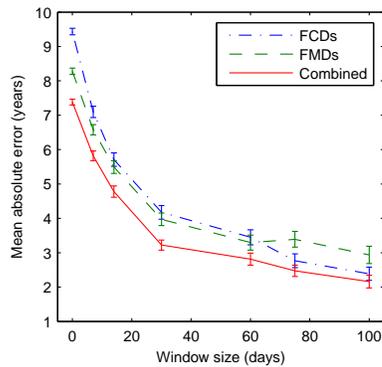


Fig. 9. Song year prediction accuracy obtained using windowed descriptors, in response to window size. Error bars denote standard errors.

synchronous representations, which yielded comparatively small gains in prediction accuracy, when combined with original frame-based features. This result suggests that for our chosen domain, temporal structure at short time scales is more advantageous, compared to temporal structure at the metrical level. One possible explanation for this behaviour is that an abundance of observations is beneficial when estimating compression rates. Alternatively, for our chosen tasks similarity judgements might predominantly be based on short-term timbral characteristics, rather than long-term structures such as motifs and chord progressions. For future work, we aim to examine in closer detail the utility of representing features at multiple time scales, and to characterise the feature spaces relevant for similarity judgements.

For similarity rating prediction, note that by biasing towards tracks with proximate chart entry dates, we attempt to control for historical changes in audio production. For song year prediction, where we do not control in the described manner, audio production may confound the association between musical similarity and chart entry date. We acknowledge that in both cases, audio production may confound the association between similarity measures and respective outcome variables, as observed in [60]. For future work, we aim to measure the degree of confounding by introducing suitable audio degradations [78]. A further issue concerns the practical impact of predicted similarity in music information retrieval. We aim to evaluate our descriptors for search, navigation and recommendation tasks, using collections of various scales.

Finally, the present work considers only a single sequential complexity measure, estimated using a single algorithm. It is conceivable that using multiple compression algorithms may reduce the error variance of estimated sequential complexity. Using alternative classification tasks, we aim to evaluate whether multiple compressors yield an improvement in prediction accuracy.

VI. ACKNOWLEDGEMENTS

This work benefited from advice and comments from Andrew J. R. Simpson, Dan Stowell, Anssi Klapuri, Mark D. Plumbley, and Armand Leroi.

REFERENCES

- [1] M. 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, 2008.
- [2] Z. Fu, G. Lu, K. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303–319, 2011.
- [3] O. Celma, "Music recommendation and discovery in the long tail," Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona, Spain, 2009.
- [4] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of audio fingerprinting," *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, vol. 41, no. 3, pp. 271–284, 2005.
- [5] J. Serrà, "Identification of versions of the same musical composition by processing audio descriptions," Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona, Spain, 2011.
- [6] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, 2006.
- [7] Y. Kim, E. Schmidt, R. Migneco, B. Morton, P. Richardson, J. Scott, J. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. 11th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2010, pp. 255–266.
- [8] 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," *The Journal of the Acoustical Society of America*, vol. 122, p. 881, 2007.
- [9] F. Barrett, K. Grimm, R. Robins, T. Wildschut, C. Sedikides, and P. Janata, "Music-evoked nostalgia: affect, memory, and personality," *Emotion*, vol. 10, no. 3, p. 390, 2010.
- [10] T. Bertin-Mahieux, D. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. 12th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2011, pp. 591–596.
- [11] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.
- [12] T. Li and M. Ogihara, "Toward intelligent music information retrieval," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 564–574, 2006.
- [13] K. West, S. Cox, and P. Lamere, "Incorporating machine-learning into music similarity estimation," in *Proc. 1st ACM workshop on Audio and Music Computing Multimedia*, 2006, pp. 89–96.
- [14] M. Slaney, K. Weinberger, and W. White, "Learning a metric for music similarity," in *Proc. 9th Intern. Conf. Music Information Retrieval (ISMIR)*, 2008, pp. 313–318.
- [15] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *Proc. Intern. Conf. Multimedia and Expo. (ICME)*, 2001.
- [16] J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?" in *Proc. 3rd Intern. Conf. Music Information Retrieval (ISMIR)*, 2002.
- [17] A. Berenzweig, B. Logan, D. 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.
- [18] M. Mandel and D. Ellis, "Song-level features and support vector machines for music classification," in *Proc. 6th Intern. Conf. Music Information Retrieval (ISMIR)*, 2005, pp. 594–599.
- [19] F. Vignoli and S. Pauws, "A music retrieval system based on user driven similarity and its evaluation," in *Proc. 6th Intern. Conf. Music Information Retrieval (ISMIR)*, 2005, pp. 272–279.
- [20] M. Levy and M. Sandler, "Lightweight measures for timbral similarity of musical audio," in *Proc. 1st ACM workshop on Audio and music computing multimedia*, 2006, pp. 27–36.
- [21] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychological Research*, vol. 58, no. 3, pp. 177–192, 1995.
- [22] K. Seyerlehner, G. Widmer, and T. Pohle, "Fusing block-level features for music similarity estimation," in *Proc. 13th Int. Conf. on Digital Audio Effects (DAFx-10)*, 2010, pp. 225–232.
- [23] F. Mörchen, A. Ultsch, M. Thies, and I. Lohken, "Modeling timbre distance with temporal statistics from polyphonic music," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 81–90, 2006.
- [24] P. Hamel, S. Lemieux, Y. Bengio, and D. Eck, "Temporal pooling and multiscale learning for automatic annotation and ranking of music audio," in *Proc. 12th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2011, pp. 729–734.

- [25] J. Wülfing and M. Riedmiller, "Unsupervised learning of local features for music classification," in *Proc. 13th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2012, pp. 139–144.
- [26] E. Pampalk, "Computational models of music similarity and their application in music information retrieval," Ph.D. dissertation, Vienna University of Technology, Vienna, Austria, 2006.
- [27] C. Lee, J. Shih, K. Yu, and H. Lin, "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 670–682, 2009.
- [28] A. Meng, P. Ahrendt, J. Larsen, and L. Hansen, "Temporal feature integration for music genre classification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1654–1664, 2007.
- [29] E. Coviello, Y. Vaizman, A. Chan, and G. Lanckriet, "Multivariate autoregressive mixture models for music auto-tagging," in *Proc. 13th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2012, pp. 547–552.
- [30] M. Li and R. Sleep, "Genre classification via an LZ78-based string kernel," in *Proc. 6th Intern. Conf. Music Information Retrieval (ISMIR)*, 2005, pp. 252–259.
- [31] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Information Theory*, vol. 24, no. 5, pp. 530–536, 1978.
- [32] J. Reed and C. Lee, "On the importance of modeling temporal information in music tag annotation," in *Proc. IEEE Intern. Conf. Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2009, pp. 1873–1876.
- [33] T. Langlois and G. Marques, "A music classification method based on timbral features," in *Proc. 10th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2009, pp. 81–86.
- [34] J. Ren and J. Jang, "Discovering time-constrained sequential patterns for music genre classification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1134–1144, 2012.
- [35] R. Foucard, S. Essid, M. Lagrange, and G. Richard, "Multi-scale temporal fusion by boosting for music classification," in *Proc. 12th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2011, pp. 663–668.
- [36] S. Dieleman and B. Schrauwen, "Multiscale approaches to music audio feature learning," in *Proc. 14th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2013, pp. 3–8.
- [37] M. Mauch and M. Levy, "Structural change on multiple time scales as a correlate of musical complexity," in *Proc. 12th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2011, pp. 489–494.
- [38] S. Dieleman, P. Brakel, and B. Schrauwen, "Audio-based music classification with a pretrained convolutional network," in *Proc. 12th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2011, pp. 669–674.
- [39] M. Slaney and M. Casey, "Locality-sensitive hashing for finding nearest neighbors," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 128–131, 2008.
- [40] C. Rhodes, T. Crawford, M. Casey, and M. d'Inverno, "Investigating music collections at different scales with AudioDB," *Journal of New Music Research*, vol. 39, no. 4, pp. 337–348, 2010.
- [41] J. Schlüter, "Learning binary codes for efficient large-scale music similarity search," in *Proc. 14th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2013, pp. 581–586.
- [42] R. Cilibrasi, P. M. B. Vitányi, and R. Wolf, "Algorithmic clustering of music based on string compression," *Computer Music Journal*, vol. 28, no. 4, pp. 49–67, 2004.
- [43] S. Dubnov, "Unified view of prediction and repetition structure in audio signals with application to interest point detection," *IEEE Trans. Audio, Speech and Language Process.*, vol. 16, no. 2, pp. 327–337, 2008.
- [44] S. Abdallah and M. Plumbley, "Information dynamics: Patterns of expectation and surprise in the perception of music," *Connection Science*, vol. 21, no. 2-3, pp. 89–117, 2009.
- [45] R. James, C. Ellison, and J. Crutchfield, "Anatomy of a bit: Information in a time series observation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 21, no. 3, p. 037109, 2011.
- [46] S. Streich, "Automatic characterization of music complexity: a multifaceted approach," Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona, Spain, 2006.
- [47] O. Lartillot and P. Toivainen, "A Matlab toolbox for musical feature extraction from audio," in *Proc. Intern. Conf. Digital Audio Effects (DAFx)*, 2007, pp. 237–244.
- [48] D. P. W. Ellis and G. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proc. IEEE Intern. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, 2007, pp. 1429–1432.
- [49] R. Plomp and W. Levelt, "Tonal consonance and critical bandwidth," *Journal of the Acoustical Society of America*, vol. 38, p. 548, 1965.
- [50] K. Jensen, "Timbre models of musical sounds," Ph.D. dissertation, University of Copenhagen, Denmark, 1999.
- [51] M. Slaney, "Auditory toolbox version 2," Interval Research Corporation, Tech. Rep., 1998.
- [52] P. Masri, "Computer modelling of sound for transformation and synthesis of musical signals," Ph.D. dissertation, University of Bristol, United Kingdom, 1996.
- [53] E. Gómez, "Tonal description of music audio signals," Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona, Spain, 2006.
- [54] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proc. 1st ACM workshop on Audio and music computing multimedia*. ACM, 2006, pp. 21–26.
- [55] R. Begleiter, R. El-Yaniv, and G. Yona, "On prediction using variable order Markov models," *Journal of Artificial Intelligence Research*, vol. 22, pp. 385–421, 2004.
- [56] J. Pressing, "Cognitive complexity and the structure of musical patterns," in *Proc. 4th Conf. Australasian Cognitive Science Society*, 1999.
- [57] G. A. Wiggins, D. Müllensiefen, and M. Pearce, "On the non-existence of music: Why music theory is a figment of the imagination," *Musicae Scientiae*, vol. 14, no. 1, pp. 231–255, 2010.
- [58] N. Goodman, "Seven strictures on similarity," in *Problems and Projects*. Indianapolis: Bobbs-Merrill, 1972.
- [59] A. Tversky, "Features of similarity," *Psychological review*, vol. 84, no. 4, p. 327, 1977.
- [60] B. Sturm, "Two systems for automatic music genre recognition: What are they really recognizing?" in *Proc. 2nd ACM Intern. workshop on Music information retrieval with user-centered and multimodal strategies*, 2012, pp. 69–74.
- [61] D. Bogdanov, J. Serrà, N. Wack, P. Herrera, and X. Serra, "Unifying low-level and high-level music similarity measures," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 687–701, 2011.
- [62] J. Downie, A. F. Ehmann, M. Bay, and M. C. Jones, "The music information retrieval evaluation exchange: Some observations and insights," in *Advances in music information retrieval*. Springer, 2010, pp. 93–115.
- [63] J. Ollen, "A criterion-related validity test of selected indicators of musical sophistication using expert ratings," Ph.D. dissertation, Ohio State University, United States of America, 2006.
- [64] B. Efron, *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982, vol. 38.
- [65] J. Serrà, H. Kantz, X. Serra, and R. Andrzejak, "Predictability of music descriptor time series and its application to cover song detection," *IEEE Trans. Audio, Speech and Language Process.*, vol. 20, no. 2, pp. 514–525, 2012.
- [66] F. Takens, "Detecting strange attractors in turbulence," *Dynamical Systems and Turbulence*, pp. 366–381, 1981.
- [67] P. Foster, S. Dixon, and A. Klapuri, "Identification of cover songs using information theoretic measures of similarity," in *Proc. IEEE Intern. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, 2013.
- [68] J. Cardoso and R. Sousa, "Measuring the performance of ordinal classification," *Intern. Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 08, pp. 1173–1195, 2011.
- [69] J. Pinto da Costa, H. Alonso, and J. Cardoso, "The unimodal model for the classification of ordinal data," *Neural Networks*, vol. 21, no. 1, pp. 78–91, 2008.
- [70] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
- [71] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.
- [72] A. Flexer and D. Schnitzer, "Effects of album and artist filters in audio similarity computed for very large music databases," *Computer Music Journal*, vol. 34, no. 3, pp. 20–28, 2010.
- [73] G. Kirchgassner, J. Wolters, and U. Hassler, *Introduction to modern time series analysis*. Springer, 2012.
- [74] G. Box, G. Jenkins, and G. Reinsel, *Time series analysis: forecasting and control*. Wiley, 2013.
- [75] C. W. Granger and R. Joyeux, "An introduction to long-memory time series models and fractional differencing," *Journal of Time Series Analysis*, vol. 1, no. 1, pp. 15–29, 1980.
- [76] M. Effros, "PPM performance with BWT complexity: A new method for lossless data compression," in *Proc. Data Compression Conf.*, 2000, pp. 203–212.

- [77] P. Hamel, Y. Bengio, and D. Eck, "Building musically-relevant audio features through multiple timescale representations," in *Proc. 13th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2012, pp. 553–558.
- [78] M. Mauch and S. Ewert, "The audio degradation toolbox and its application to robustness evaluation," in *Proc. 14th Intern. Society for Music Information Retrieval Conf. (ISMIR)*, 2013, pp. 83–88.