Information-Theoretic Measures of Predictability for Music Content Analysis

Peter Foster

Submitted to the University of London in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Centre for Digital Music

School of Electronic Engineering and Computer Science

Queen Mary University of London

November, 2014

I, Peter Foster, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature:

Date: October 30, 2014

Details of collaboration and publications: All collaborations and previous publications related to this thesis are described in Section 1.4.

Abstract

This thesis is concerned with determining similarity in musical audio, for the purpose of applications in music content analysis. With the aim of determining similarity, we consider the problem of representing temporal structure in music. To represent temporal structure, we propose to compute information-theoretic measures of predictability in sequences. We apply our measures to track-wise representations obtained from musical audio; thereafter we consider the obtained measures predictors of musical similarity. We demonstrate that our approach benefits music content analysis tasks based on musical similarity.

For the intermediate-specificity task of cover song identification, we compare contrasting discrete-valued and continuous-valued measures of pairwise predictability between sequences. In the discrete case, we devise a method for computing the normalised compression distance (NCD) which accounts for correlation between sequences. We observe that our measure improves average performance over NCD, for sequential compression algorithms. In the continuous case, we propose to compute information-based measures as statistics of the prediction error between sequences. Evaluated using 300 Jazz standards and using the Million Song Dataset, we observe that continuous-valued approaches outperform discrete-valued approaches. Further, we demonstrate that continuous-valued measures of predictability may be combined to improve performance with respect to baseline approaches. Using a filter-and-refine approach, we demonstrate state-of-the-art performance using the Million Song Dataset.

For the low-specificity tasks of similarity rating prediction and song year prediction, we propose descriptors based on computing track-wise compression rates of quantised audio features, using multiple temporal resolutions and quantisation granularities. We evaluate our descriptors using a dataset of 15 500 track excerpts of Western popular music, for which we have 7 800 web-sourced pairwise similarity 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.

Contents

List of Figures

List of Tables

Acknowledgements

I am indebted to a great number of individuals who have helped me to undertake PhD research.

First and foremost, I sincerely thank my PhD supervisors Simon Dixon and Anssi Klapuri for their advice and guidance over the course of my research. I sincerely thank Mark Plumbley for additional advice and guidance, and for welcoming me at the Centre for Digital Music at Queen Mary. I have enjoyed tremendously undertaking my PhD at the Centre for Digital Music and at Queen Mary.

In addition to my supervisors, I am grateful to Matthias Mauch, Dan Stowell and Andrew Simpson for providing detailed feedback on the work described in this thesis.

I am further grateful to Matthias Mauch for making available data without which this work would not have been possible. Further, I am grateful to Dan Ellis and Graham Poliner; Ron Begleiter; Olivier Lartillot and Petri Toiviainen; Bertin-Mahieux et al.; Qian et al. for having made available their code and data prior to this work. I am also grateful to the IT staff at the School of Electronic Engineering and Computer Science for their excellent computing facilities.

I thank the following individuals for their kind help and advice, and support as friends and colleagues: Samer Abdallah, Jakob Abeßer, Amélie Anglade, Thierry Bertin-Mahieux, Daniele Barchiesi, Mathieu Barthet, Chris Baume, Emmanouil Benetos, Chris Cannam, Tian Cheng, Magdalena Chudy, Alice Clifford, Henrik Ekeus, Sebastian Ewert, György Fazekas, Luis Figueira, Dimitrios Giannoulis, Aris Gretsistas, Steven Hargreaves, Chris Harte, Ho Huen, Maria Jafari, Maksim Khadkevich, Holger Kirchhoff, Şefki Kolozali, Katerina Kosta, Pamela Lawson, Armand Leroi, Shengchen Li, Robert Macrae, Boris Mailhé, Martin Morrell, Tim Murray-Browne, Ken O'Hanlon, Juan Bello, Marcus Pearce, Colin Powell, Elio Quinton, Suman Ravuri, Andrew Robertson, David Ronan, Siddharth Sigtia, Jordan Smith, Janis Sokolovskis, Chunyang Song, Yading Song, Adam Stark, Michael Terrell, Mi Tian, Robert Tubb, Bogdan Vera, Siying Wang, Steve Welburn, Geraint Wiggins, Sonia Wilkie, Luwei Yang, Melissa Yeo, Asterios Zacharakis.

Finally, I thank all my family and Renata Sadibolova for their unwavering moral support.

This work was supported by a UK EPSRC DTA studentship.

Chapter 1

Introduction

The proliferation of music in digital formats poses considerable challenges and opportunities for organising and navigating collections of recorded music. Coinciding with the growth and increasing ubiquity of internet services since the new millennium, online music services are replacing physical formats such as CDs as a mechanism for disseminating recorded music. Online services allow users to purchase audio tracks for download, akin to a conventional music store. Alternatively, music may be streamed in real-time, akin to a conventional radio station. However, in both such cases, users are offered near-instantaneous and personalised access to large collections: the online music store iTunes reports a collection of more than 40 million tracks[1], whereas the streaming service Spotify reports a collection of more than 20 million tracks[2]. For download and streaming services, the Recording Industry Association of America reports respective revenues of $1305 million and $859 million, for the first half of 2014 (Friedlander, 2014). For the same period, downloads and streaming account for 41% and 27% of all revenues, compared to 28% of all revenues for physical formats. These revenue statistics suggest that compared to the pre-internet age, the dissemination of recorded music has been transformed.

The field of music information retrieval (MIR) is concerned with investigating methods for providing access to music collections (Casey et al., 2008b). Downie (2003) distinguishes between a number of musical facets, including pitch, rhythm, harmony, timbre, lyrics, performance instructions and the bibliographic facet of musical works. In MIR, it is sought to obtain repre-

[1] `http://www.apple.com/itunes/`, retrieved October 2014.
[2] `https://press.spotify.com/uk/information/`, retrieved October 2014.

sentations for such facets that allow music to be organised for subsequent navigation by a user. On the basis of music representations, we may distinguish among tasks spanning audio fingerprinting, genre classification, artist classification, mood classification, version identification, and structural segmentation (cf. Cano et al., 2005; Fu et al., 2011b; Casey et al., 2008b; Kim et al., 2010; Serrà, 2011; Paulus et al., 2010). With a view to performing such tasks, we may consider disparate sources of information: the sampled audio signal, musical score representations, or annotations including textual meta-data. Focussing on the sampled audio signal, *music content analysis* investigates methods which may serve as alternatives to manual annotation processes, when the latter are infeasible, unavailable or amenable to be supplemented (Casey et al., 2008b; Celma, 2009). This thesis relates to music content analysis.

This chapter motivates and summarises the work described in this thesis. Section 1.2 discusses the research questions addressed in this thesis. Further, Section 1.3 summarises novel contributions. Section 1.4 lists related publications by the author. Finally, Section 1.5 briefly outlines the following chapters.

1.1 Music Content Analysis

A variety of users would benefit from the application of music content analysis in MIR systems. Music consumers would benefit from novel recommendation systems enabling them to search and discover music, in turn benefiting the music industry. Besides consumers, individuals such as musicologists, artists, cataloguers, music tutors and therapists would benefit from similar systems. For the particular cases of audio fingerprinting, artist classification and version identification, record labels may seek to identify instances of copyright infringement; cataloguers may seek to identify instances of misattribution in musical works. Music content analysis has found widespread commercial application in the online service provided by EchoNest[3]. For the purpose of scientific inquiry, music content analysis has the potential for cross-disciplinary investigations in musicology and music psychology (Serra et al., 2013).

Figure 1.1 provides a schematic summary of stages involved in a typical music content analysis system. Following feature extraction, track-wise representations of feature sequences are obtained. Such representations may then be used to classify tracks. Alternatively, distances between representations may be computed to obtain a measure of pairwise similarity between

[3] `http://the.echonest.com/`, retrieved October 2014.

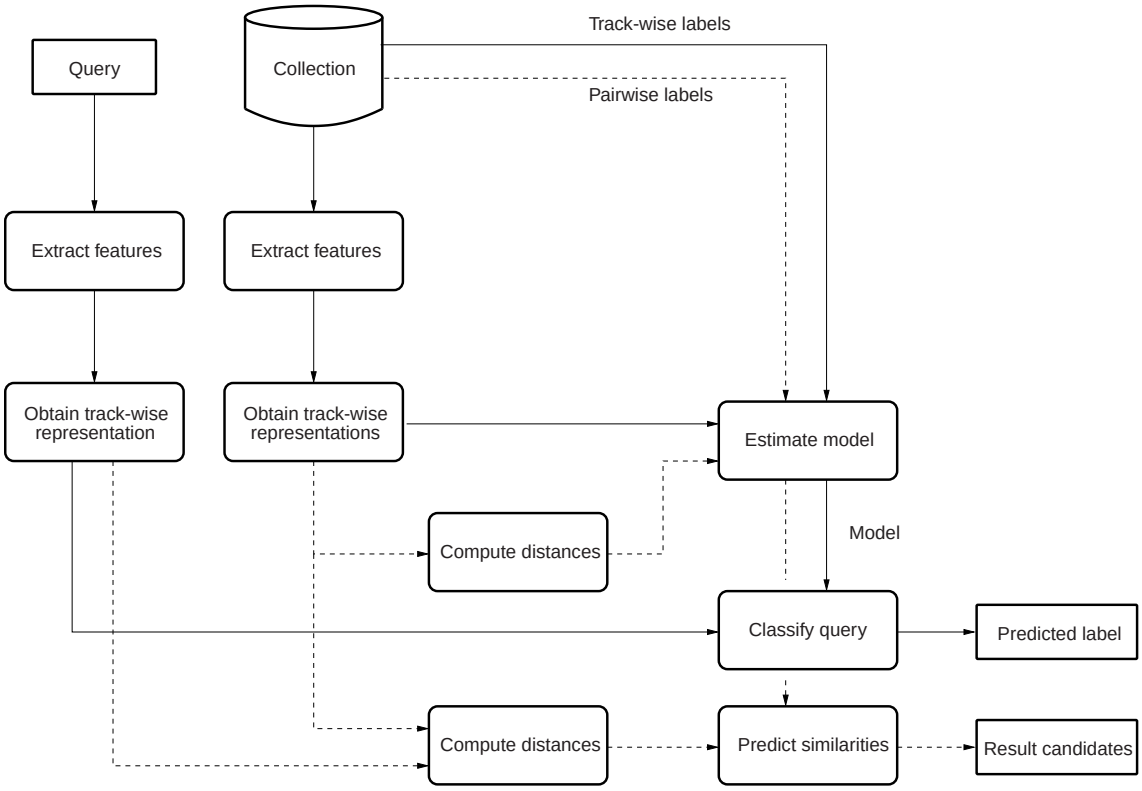


Figure 1.1: Stages in a typical music content analysis system. The system incorporates a classifier which is subsequently used to classify tracks. Dashed lines indicate stages for an alternative 'query-by-example' system.

tracks. The second approach may be used to retrieve similar tracks in the collection, with respect to the query ('query-by-example').

Classification tasks in music content analysis typically perform a mapping of audio tracks to labels. For example, in audio fingerprinting we seek a mapping under which labels are unique to particular recordings of a piece of music. In version identification, we seek a mapping under which labels are unique to a piece of music for which there may exist multiple versions. Similarly, in genre and mood classification, labels are ideally unique to genres and moods, respectively. If we consider the equivalence classes with regard to the described mappings, the aforementioned fingerprinting, version identification, genre classification, mood identification tasks relate to similarity measures between audio tracks. Thus, how to determine musical similarity is a central problem in music content analysis.

We may further characterise the described tasks, by considering the typical size of equivalence classes in relation to the size of the collection. In typical applications of audio fingerprinting such as copyright infringement detection or recording identification, given a query the set of rele-

vant tracks is small in relation to the collection. In genre and mood identification for applications such as music recommendation or playlist generation, since there may exist many exemplars for a given genre or mood, the set of tracks relevant to the query may be large in relation to the collection. Thus, we may distinguish between music content analysis tasks according to the degree of *specificity* associated with tracks deemed relevant to a query (Byrd, 2007; Casey et al., 2008a). We collectively refer to tasks such as genre and mood classification as *low-specificity tasks*. Note that version identification is deemed to have mid-level, diffuse specificity in comparison to audio fingerprinting and genre classification, since versions may differ from the original song to varying degrees, potentially involving various musical facets. For example, version identification on the one hand incorporates identifying remasters of a recording, where remasters may differ only in musical timbre and noise characteristics. On the other hand, version identification incorporates identifying *cover versions* of a song, where cover versions may additionally differ in timbre, tempo, timing, structure, key, harmony, or lyrics (Serrà, 2011).

As discussed by Rhodes et al. (2010), the described tasks may involve collections at various scales. Fingerprinting applications typically require large collections of approximately 10^6 tracks to be of practical use, due to the task's high specificity. In contrast, version identification applications exist for both for small collections of approximately 10^3 tracks, in addition to large collections. Particular to large collections is the requirement of computational efficiency, since using a linear scan of tracks may be prohibitively slow to determine similarity with respect to a query. One possible approach to attaining computational efficiency involves using scalable algorithms with sub-linear time complexity (cf. Slaney and Casey, 2008).

1.2 Research Questions

RQ1: How may we obtain representations of temporal structure in music?

From the perspective of music content analysis, the fundamental research question which this thesis addresses is how to obtain representations of temporal structure in music. This question is of considerable interest, since music is intrinsically a temporal phenomenon. As we review in Chapter 2, among musical similarity tasks of diverse specificities, we may distinguish between diverse methods for obtaining sequences of features from musical audio and subsequently performing similarity comparisons using such sequences.

In version identification (cf. Section 2.4.2), existing approaches are typically based on ex-

tracting sequences of mid-level features representing harmonic content in musical audio. Given a collection of tracks, among which we seek to identify those tracks corresponding to versions of a query track, pairwise similarities are typically determined either by performing sequence alignment or by comparing models of feature sequences. We may distinguish between methods based on strings, and methods based on sequences of continuous-valued audio features. As evidenced by recent investigations which employ either discrete-valued or continuous-valued representations, accurate as well as computationally efficient version identification remains a challenge. We perform investigations on cover song identification with the aim of improving beyond the state of the art in accuracy, where we consider both discrete-valued and continuous-valued approaches. We address the challenge of computationally efficient retrieval in our evaluation, by considering a dataset on the scale of 10^6 tracks. We define a cover song as a rendition of a previously recorded piece of music, particular to the practice in Western popular music where renditions may be created by different artists (cf. Serrà, 2011).

In low-specificity similarity tasks (cf. Section 2.4.3), existing approaches typically involve extracting sequences of either low-level or mid-level features. Thereafter, we may distinguish between methods which disregard the temporal order of features (thus discarding information on temporal structure), and methods in which temporal order of features is retained. The former so-called 'bag-of-features' approach involves estimating distributions of individual observations in a feature sequence. The bag-of-features approach may be used to obtain relatively low-dimensional, duration-invariant representations of features in a given track, subsequently allowing computationally efficient methods to be used for determining pairwise similarity. Whereas bag-of-features approaches have been widely applied, their relative convenience contrasts with the disadvantage inherent in discarding temporal information. As evidenced by recent investigations which use bag-of-features representations, low-specificity similarity tasks remain a challenge, from the perspective of accurate as well as computationally efficient retrieval. The investigations described in this thesis address the question of how to preserve temporal information in features, in a manner which permits computationally efficient retrieval.

RQ2: Is it possible to use information-theoretic measures of predictability to represent temporal structure in music?

Relating to **RQ1**, this thesis specifically investigates the use of information-theoretic measures of predictability as a representation of temporal structure in music. As discussed in Chapter 2, our

approach is inspired by music-psychological models of expectation which are based on score representations of music. Among such approaches, statistical models of sequences are constructed; Shannon information is then used to quantify predictive uncertainty. We note similar methods based on Shannon information which have been proposed to quantify predictability in musical audio, however these approaches to date have not been evaluated quantitatively in music content analysis tasks. We motivate our approach on the basis of these considerations.

RQ3: To what extent can information-theoretic measures of predictability be used to determine musical similarity?

Described in chapters 4 and 5, the investigations reported in this thesis are concerned with cover song identification, similarity rating prediction and song year prediction, all of which are musical similarity tasks. In applying our approach to these tasks, we seek to establish whether information-theoretic measures of predictability may be used to determine musical similarity.

Our measures of predictability are designed to capture temporal structure in sequences, thus we hypothesise that our approach captures temporal structure in music. From evidence in support of using our measures to determine musical similarity, we may conclude that our measures indeed capture temporal structure in music. Thus, from our observations we may subsequently arrive at conclusions concerning **RQ1** and **RQ2**.

Our investigations contrast two means of computing predictive uncertainty. For cover song identification, we quantify predictive uncertainty of one feature sequence relative to another sequence, which we refer to as *pairwise predictability*. We use pairwise predictabilities as similarities between tracks. This contrasts with our approach to similarity rating prediction and song year prediction, where we quantify predictive uncertainty in a single feature sequence. To determine similarities, in the latter case we consider a metric space on track-wise predictabilities. That is, we deem two tracks similar, if their feature sequences yield similar amounts of predictive uncertainty. In Section 2.3.4 we discuss the suitability of pairwise versus track-wise predictability for the considered similarity tasks, based on their respective specificities.

In both cases, we quantify the performance of our measures by computing agreement between predicted and annotated similarity data: for cover song identification, our predictions are tracks ranked with respect to query tracks; our annotations are the sets of tracks which are cover songs of query tracks. For similarity rating prediction, our predictions and annotations are pairwise similarity ratings between tracks. For song year prediction, our predictions and annotations are

the chart entry dates of tracks.

RQ4: Which measures of predictability are useful for determining musical similarity?

As previously described in **RQ1**, we contrast discrete-valued and continuous-valued representations in our investigations on cover song identification. As we discuss in Chapter 3, Shannon information allows us to express pairwise similarity between sequences in a number of ways. Thus, to establish which measures of predictability are useful in determining musical similarity, we compare a variety of possible measures. Notably, we may interpret the normalised compression distance (NCD) as a further information-theoretic approach, which has previously been applied in cover song identification. Our evaluation aims to establish the performance of alternative distance measures to the NCD. As we detail in Chapter 4, next to the behaviour of information-theoretic measures, we compare in detail the effect of interchanging techniques used to estimate them. Such comparisons consider the choice of statistical model applied to strings. Further, in relation to evaluated discrete-valued measures of predictability we contrast and propose analogous continuous-valued measures of predictability which do not require feature quantisation and thus may be applied directly to feature sequences.

RQ5: Which feature representations are useful for determining musical similarity?

Whereas our investigations on cover song identification are based on a single mid-level feature, in our investigations on similarity rating prediction and song year prediction we evaluate a set of low-level and mid-level features. For each feature, we evaluate the utility of quantifying predictive uncertainty relative to a 'bag-of-features' approach, where we downsample each feature sequence using a set of specified rates. In this manner, we seek to establish whether representations at multiple time scales are useful for determining similarity. We motivate this research question on the basis that diverse feature representations have been proposed for low-specificity tasks; to date the potential of representing features at multiple time scales has not been investigated widely.

RQ6: How may we quantify similarity between sequences?

While this thesis is primarily concerned with applications in music content analysis, our approach relates to the general problem of quantifying similarity between sequences. As described in Chapter 4, we propose a variant of the NCD which eliminates a deficiency in the existing

approach, when applied to sequences generated by Markov information sources. In addition to evaluations involving cover song identification, we compare the theoretical behaviour of our measure using artificially generated sequences. Further, we propose methods for computing information-theoretic measures of similarity which unlike the NCD do not require feature quantisation.

RQ7: How might we perform computationally efficient retrieval?

As was previously mentioned for **RQ1**, in our investigations we address the question of how to determine similarity in a computationally efficient manner, which we view as a prerequisite for retrieval using large-scale datasets. Concerning cover song identification, while our information-theoretic measures are relatively computationally expensive, we demonstrate that they may be combined with more computationally efficient methods, in a two-stage 'filter-and-refine' process. Using this approach, we attain state-of-the-art performance using 10^6 tracks. While our evaluated filter-and-refine approach uses an exhaustive linear scan to produce an initial ranking of tracks relative to a query, the initial ranking is based on pairwise comparisons using a metric. This approach allows the potential use of sub-linear techniques in place of a linear scan. Similarly, our comparisons of track-wise measures of predictive uncertainty are based on a metric space. Thus, we allow for the possibility of using sub-linear retrieval techniques.

1.3 Novel Contributions

The novel contributions in this thesis may be summarised as follows:

- **Empirical evidence for the utility of information-theoretic measures of predictability in music content analysis.** This thesis shows that information-theoretic measures of predictability may be applied to musical similarity tasks. For the purpose of cover song identification, we compare contrasting measures of pairwise predictability, which we use to compute similarities between pairs of feature sequences. In the NCD, there exists a widely-applied information-theoretic measure of similarity which has previously been applied to cover song identification, however to date no extensive comparison has been made between it and alternative approaches. We demonstrate that our alternative measures to the NCD yield state-of-the-art accuracy; moreover we demonstrate that similarities may be combined to improve accuracy with respect to baseline approaches.

For the purpose of similarity rating prediction and song year prediction, we propose predictability as a means of summarising feature sequences across individual tracks. We demonstrate that such measures are relevant for determining similarity in similarity rating prediction and song year prediction. Moreover, we demonstrate that our measures may be combined to improve accuracy with respect to baseline bag-of-features representations.

- **Methods for quantifying similarity between sequences.** This thesis proposes novel measures of similarity between sequences. Firstly, in the normalised compression distance with alignment (NCDA), we propose a variant of NCD which accounts for correlation between discrete-valued observations generated by Markov information sources. As we observe for artificially generated sequences, NCDA better characterises distances between sequences compared to NCD. Applied to cover song identification and based on quantised audio features, NCDA improves cover song identification accuracy compared to NCD, when applied to the Lempel-Ziv (LZ) compression algorithm. Next to NCDA, we propose measures of similarity which may be computed on continuous-valued sequences, thus requiring no preceding feature quantisation step. The approach relies on combining a nearest-neighbours prediction technique with parametric distribution estimation, which in combination represents a novel contribution. We observe that continuous-valued similarity measures outperform discrete-valued approaches.

- **A method for quantifying regularity in feature sequences.** This thesis proposes a measure of temporal regularity in feature sequences. The approach is based on quantising feature sequences and subsequently determining compression rates of strings. Further, we downsample feature sequences using a set of specified rates. In this way, we quantify predictive uncertainty at multiple time scales. Exploratory and quantitative analysis reveals that the proposed measure captures temporal structure relevant for low-specificity similarity tasks.

1.4 Publications

This thesis is based on the following publications:

- Peer-reviewed:

  – P. Foster, M. Mauch, and S. Dixon. Sequential complexity as a descriptor for musical

similarity. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22 (12):1967–1977, 2014b

  - P. Foster, S. Dixon, and A. Klapuri. Identification of cover songs using information theoretic measures of similarity. In *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 739–743, 2013

- In progress:

  - P. Foster, A. Klapuri, and S. Dixon. Identifying cover songs using information-theoretic measures of similarity. *arXiv preprint arXiv:1407.2433*, 2014a

In the listed publications, PF performed all scientific tasks from research question formulation up to and including manuscript writing. Co-authors SD, AK, MM provided advice during meetings and via electronic correspondence, in addition to comments on manuscripts. The author's main supervisor was SD. The data used in Foster et al. (2014b) were kindly provided by MM. Finally, the author is grateful to the individuals given in the Acknowledgements section of this thesis, for further advice and comments on manuscripts.

The work described in Foster et al. (2014a) substantially extends preceding work described in Foster et al. (2013). Notably, it extends the set of evaluated similarity measures; it applies the methods to an additional, large-scale dataset; it investigates combining similarity measures.

1.5 Thesis Outline

The remainder of this thesis is organised as follows:

- Chapter 2 reviews existing work on the similarity tasks considered in this thesis. Furthermore, we consider the phenomenon of musical expectation. Owing to models of musical expectation, we motivate the decision to investigate the use of information-theoretic measures of predictability for determining musical similarity.

- Chapter 3 briefly reviews information-theoretic concepts relevant to our investigations and proposed methods. Next to measures of uncertainty on sequences of discrete and continuous random variables, we review measures of disparity between sequences.

- Chapter 4 describes investigations on cover song identification. This chapter is based on

Foster et al. (2014a); it includes additional analysis of NCDA and experiments using artificially generated sequences in Section 4.2.2.

- Chapter 5 describes investigations on similarity rating prediction and song year prediction. This chapter is based on Foster et al. (2014b).
- Chapter 6 discusses the findings in this thesis, before examining possibilities for future research.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, we review and discuss concepts and related work on musical expectation and computational methods for determining similarity in musical audio.

Since music is an inherently subjective phenomenon, we consider musical concepts from the perspective of the listener. We begin in Section 2.2 by reviewing the musical phenomena on which our work is based, where we consider perceptual and cognitive processes involved in music listening. We then proceed to a discussion of musical expectation in Section 2.3. As discussed in Section 2.3.4, we identify models of musical expectation as a possible approach to determining musical similarity. Based on existing work in expectation modelling, we propose to use information-theoretic measures of predictability in our own inquiry.

In Section 2.4, we review and discuss related work on musical similarity from the perspective of music content analysis. We begin in Section 2.4.1 by reviewing audio features relevant to our work.We then describe approaches to determining musical similarity relevant to our investigations in chapters 4 and 5, which relate to cover song identification and low-specificity similarity. Finally, in Section 2.5 we summarise and conclude our discussion.

2.2 Music-Theoretical and Psychological Background

According to Patel (2010, Chap. 5), music involves perceptually discrete elements and conventions for combining such elements into sequences. Because music is a human universal (Cross,

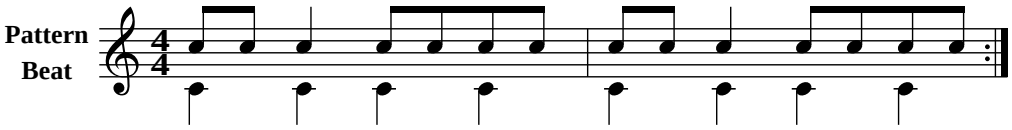


Figure 2.1: Example score with a rhythmic pattern (upper voice) and corresponding beat (lower voice).

2001), yet with differing musical practices across cultures, providing an exhaustive account of musical elements and conventions for combining them is beyond the scope of this thesis. In the following, we provide a brief description of musical phenomena, where we confine our discussion to music which relates to conventions of the common practice era (CPE), ranging from around 1600 to 1900 CE (Kennedy et al., 2013, p. 373). We discuss two important phenomena, namely pitch and rhythm. For a more detailed discussion of musical concepts introduced in the following, we refer to Kennedy et al. (2013).

2.2.1 Rhythm

Rhythm relates to the dimension of time at scales above 120ms and below 1.5s (Bolton, 1894; Repp, 2003). Fraisse (1982) defines rhythm as a perceived ordered succession of distinct events. Since events are ordered, listeners are able to predict future events based on what has been perceived. As suggested by Mach (1865), perceived rhythm relates closely to anticipatory motor activity, such as tapping. Perceptual events may be distinguished according to duration, pitch, loudness, or *timbre*. As discussed by Rasch and Plomp (1999), we define timbre as the quality which allows the listener to discern dissimilarity between steady-state tones with identical pitch and loudness.

Perceived rhythm may coincide with a sequence of periodic and identical perceptual events. This sequence of periodic events is the *beat*, also referred to as the *tactus*. The period between successive events determines the tempo, measured in beats per minute. A phenomenon related to beat is metre, which London (2012, Chap. 1) defines as the periodic anticipatory schema which enables a listener to infer beat and direct motor behaviour. From the perspective of music theory, metre determines how beats are counted, divided and grouped into bars in common practice notation (CPN). In the listener, metre influences perceived rhythmic patterns. Figure 2.1 displays an example score in CPN, where we include a repeated rhythmic pattern along with notated beat; in CPN the time signature and bar-lines provide information on metrical structure.

2.2.2 Pitch

A listener may perceive pitch produced by pitched instruments, which in contrast to percussive instruments produce approximately periodic sound waves. Pitched instruments typically produce complex tones, with partials closely approximating integer multiples of the fundamental frequency. The phenomenon of pitch relates to fundamental frequency. Listeners perceive constant ratios of frequencies as having constant difference in pitch; pitch perception is approximately linear with respect to the logarithm of frequency. For a more detailed discussion of pitch, we refer to de Cheveigne (2005). Notes have pitch and duration, thus a succession of notes incorporates rhythmic perceptual characteristics.

As summarised by Krumhansl (2000), pitch in music differs from speech in the use of musical intervals. A musical interval is the distance between two pitches, which we may quantify as the ratio of fundamental frequencies associated with the pitches. The ratio of fundamental frequencies explains listeners' perception of consonant and dissonant intervals; consonant intervals closely approximate simple ratios such as 2:1 (*octave*), 3:2 (*perfect fifth*), 4:3 (*perfect fourth*). Listeners perceive pitches which are an integer number of octaves apart as having identical *chroma* (cf. Balzano and Liesch, 1982) . From the perspective of music theory, such pitches have identical pitch class. Except for the constraint of hearing range, pitch perception is therefore cyclic. In CPE music, there are 12 pitch classes; the associated chromatic scale consists of pitches whose precise intervals are determined by the tuning system used. While there exist diverse tuning systems for CPE music, we may consider twelve-tone equal temperament the most widespread from the mid-nineteenth century onwards (cf. Kennedy et al., 2013, p. 848). Equal temperament specifies a common interval ratio of $\sqrt[12]{2}$ between adjacent pitches (*semitones*) in the chromatic scale. Table 2.1 lists pitch classes and associated musical symbols. Assuming equal temperament, the set of possible fundamental frequencies $\mathcal{F}_p$ for a given pitch class p is given as

$$\mathcal{F}_p = \{2^{p/12+n} f : n \in \mathbb{Z}\} \tag{2.1}$$

where f denotes the pitch reference, by convention often $f = 440$Hz. Based on a hearing range of 16Hz to 16kHz, humans may distinguish among 120 possible pitches in the equal-tempered scale (Olson, 1967).

From the perspective of music theory, music from the CPE is based on *diatonic* scales, with each scale based on a sequence of seven distinct pitch classes from the chromatic scale. Pitches

p	0	1	2	3	4	5	6	7	8	9	10	11
Musical symbols	A	A♯ B♭	B	C	C♯ D♭	D	D♯ E♭	E	F	F♯ G♭	G	G♯ A♭

Table 2.1: Pitch classes p and associated musical symbols.

in diatonic scales are enumerated as *degrees*. A scale is associated with a tonal centre (the *tonic*), which is the pitch class of the first degree. In addition, a scale is associated with a particular *mode*, which determines the intervals between successive degrees. In CPE music, two modes predominate, namely major and minor modes. Scales constrain the set of pitch classes used in melody (pitch succession) and harmony (simultaneous production of pitch based on chords). In CPE music, chord progressions are described in terms of scale degrees associated with individual chords. Musical *key* is determined by the tonic and mode.

2.2.3 Musical Structure and Cognition

In CPE music, both pitch and rhythm are organised hierarchically at multiple time scales. From the perspective of music theory, notes and note onsets respectively form distinct motifs, which in turn form phrases. Listeners may perceive boundaries between such units and may infer relations between units involving repetition, variation and contrast. For the case of harmony, while chord changes typically occur at time scales above the motivic level of rhythm and melody, analogously listeners may infer hierarchical repetition, variation and contrast within chord progressions. In the present discussion, repetition, variation and contrast are similarity relations. Therefore, the concept of similarity is fundamental to music listening.

Music listening involves subjective judgement using information not readily represented in the musical score: we distinguish between fundamental perceptual events such as notes and chords, and the similarity relations inferred based on sequences of perceptual events. Whereas notes and chords are directly represented in CPN, musical structure is not represented in general. Following Wiggins (2007), we refer to the mental processes involved in identifying musical structure as cognitive processes. Such processes contrast with perceptual processes which relate to individual notes and chords. Conceptually, perceptual processes relate to the *musical surface* (Jackendoff, 1987), whereas cognitive processes operate below the musical surface.

From the perspective of engineering, we adopt the view that performance gains in a music content analysis system might be obtained by emulating perceptual or cognitive processes. Focussing on cognitive expectation processes involved in music listening, in the following section

we provide a brief discussion of such processes. We stress that we do not aim at simulating expectation processes, rather we motivate our discussion to provide sufficient background for later comparison with our own approach, which is inspired by existing models of musical expectation.

2.3 Musical Expectation

An account of the importance of expectation in music listening is provided by Hanslick (1891 / 1986), who posits that the fulfilment or violation of expectations influences the listener's affective response. This view is shared by the music theorist Meyer (1956), who draws attention to the notion that expectations may be revised as musical events unfold. According to these two perspectives, elucidating the operation of listening processes involves examining the objective (and thus intrinsic) capacity for music to elicit expectations. In turn, the listener forms subjective expectations via an evolving internal model of what has been perceived. As summarised in Huron (2006), early investigations involving isolated non-musical stimuli reveal that the reaction time associated with identifying a stimulus is proportional to the information-theoretic entropy of observable stimuli (Hick, 1952; Hyman, 1953). Thus, one might posit that the statistical properties of musical stimuli are relevant in music listening. We discuss illustrative studies separately for rhythm and pitch facets. Following Cont (2008, p. 13), we define expectation as a mental representation which is coupled with subsequent prediction.

2.3.1 Pitch

Following the work of Hick, Hyman, a number of studies examine listeners' responses to isolated pitch stimuli. Greenberg and Larkin (1968) assess listeners' ability to detect pitched probe signals embedded within a noise signal, observing that identification accuracy improves for stimuli which are close in pitch. Using a *probe-tone* paradigm, Howard et al. (1984) examine listeners' ability to detect pitched probe signals in a sequence. The authors observe that listeners more accurately identify omitted probe signals when the entire sequence consists of identically pitched events. We may interpret both results as evidence that listeners conditionally direct attention in the pitch dimension during listening, based on the immediate past. Attention might be considered the result of expectation: in the presence of predictable stimuli, expectation enables more accurate perception than is otherwise achieved for unpredictable stimuli (Huron, 2006, p. 43).

Expectation suggests a memory component, since it requires reference to what has previ-

ously occurred—evidently we are able to form accurate predictions when attending to a familiar melody. Whereas in the work by Howard et al. (1984) the memory component might be influenced by auditory sensory memory alone, subsequent studies examine the possible influence of long-term memory. Carlsen et al. (1970) observe that when presenting listeners with a melody, sung melodic continuations vary across cultures. Curtis and Bharucha (2009) observe an effect similar to Greenberg and Larkin (1968), whereby probe-tones are more readily misidentified as being present in a preceding sequence if the probe-tone is obtained from a culturally familiar scale, compared to the case where the probe-tone is obtained from an unfamiliar scale.

That music listening involves statistical learning is evidenced by the work of Saffran et al. (1999). The authors generate pitch sequences constructed from three-note figures, such that two disjoint sets of three-note figures contribute to two classes of pitch sequence. Both three-note figures and resulting pitch sequences are isochronous. Listeners are exposed to a pitch sequence from one class, in a learning phase. In a testing phase, listeners are asked to judge pairs of three-note figures according to familiarity. Pairs are constructed so that one of the constituent figures is sampled from set to which they were previously exposed, by way of the pitch sequence. Thus, each pair contains a familiar and an unfamiliar figure. Saffran et al. (1999) observe that listeners judge figures to which they were previously exposed as more familiar than those figures to which they were not exposed. The authors' conclusion that listeners learn statistical properties of tone sequences, relates to the fact that there exist no cues on boundaries between figures. Specifically, one may conclude that listeners implicitly learn probabilities of pitch transitions.

One might further ask whether expectation and statistical learning are involved in the cognition of scale degrees. Recall from Section 2.2.2 that scale degree is an abstraction of pitch. To establish whether listeners form expectations based on scale degree, Aarden (2003, chap. 4) exposes listeners to melodies, where listeners are asked to indicate for each successive note the intervallic direction with respect to immediately preceding notes. The reaction time between a note onset and the ensuing rating is taken as a measure of expectancy, with fast reaction times indicating a strong expectancy. Based on an analysis of folk melodies in major keys, Aarden observes in a multivariate analysis incorporating melodic context that scale degree probability predicts reaction time. Thus, the obtained results suggest that listeners incorporate both melodic context and scale degree when forming expectations and that scale degree distributions are implicitly learnt by listeners.

Expanding the role of expectation observed by Aarden, Huron (2006, p. 153) proposes that expectation accounts for the subjective experiences (*qualia*) associated with scale degrees. Huron first observes quantitatively that scale-degree patterns evoking a sense of closure in the listener (*cadences*) have relatively high statistical regularity compared to non-cadential patterns, as quantified using probabilities of degree transitions. Observing that the end of a cadence entails a relatively weak statistical association with successive scale degrees, Huron posits that a relative absence of expectation in the listener about successive scale degrees contributes to the sense of closure experienced following a cadence. Based on an exploratory analysis of descriptive terms associated with scale degree qualia, descriptive categories relating to certainty, tendency, completion, correlate with transition probabilities between scale degrees. Thus, listeners appear to be sensitive to transition probabilities of scale degrees. The hypothesis that scale degree schemata are acquired using statistical learning, is supported by cross-cultural investigations (Eerola, 2004).

2.3.2 Rhythm

Concerning the phenomenon of rhythm, an analogous role emerges of expectation based on statistical learning. Using the method of probe-tones, Jones et al. (2002) request listeners to judge the intervallic direction of a probe tone relative to a reference tone. Interpolated between the probe tone and reference tone is a sequence of isochronous 'distractor' tones which individually vary in pitch. Thus, the distractor tones are intended to establish rhythm in the listener. The authors vary the onset time of the probe tone, observing that listeners most accurately judge intervallic direction when the onset of the probe tone occurs with the beat implied by the distractor tones. In contrast, a delayed or advanced onset of the probe tone causes rating accuracy to diminish. The authors conclude that listeners' attention is directed by temporal expectancies about future events which arise during exposure to a rhythmic pattern.

Pertaining to statistical learning in rhythm cognition, Desain et al. (2003) examine the categories of rhythmic pattern assigned by listeners to three-note patterns. To this end, the authors vary the inter-onset intervals determining patterns on a continuous scale. The obtained patterns are exposed to musicians, who are asked to transcribe each pattern individually. Plotting the range of onset intervals with respect to each category, the authors observe varying inter-onset tolerances across categories. The authors explain such tolerances by examining the empirical distribution of rhythmic patterns in a corpus of Western music, observing a correspondence between cate-

gory size and rate of occurrence. Thus, the mechanism of rhythm categorisation appears to be informed by pattern probabilities. In particular, frequently occurring rhythmic patterns are associated with comparatively large inter-onset tolerances and may absorb rarely occurring patterns, in terms of listener categorisation. Desain et al. (2003) provide a Bayesian account of rhythmic familiarity, which implicates statistical learning in the cognition of rhythmic patterns.

2.3.3 Modelling Musical Expectation

Techniques for modelling musical expectation have been proposed from music-theoretical as well as psychological and engineering domains. As noted by Rohrmeier and Koelsch (2012) we may view sampling estimates of musical event probabilities as models of musical expectation, since such statistics may be used to form predictions. As a result, a model of musical expectation is contained in straightforward scale degree transitions (Piston, 1978), as well as complex statistical models (Pearce and Wiggins, 2004).

Music-Theoretical Approaches

In the implication-realisation (I-R) model, Narmour (1990) proposes a theory of melodic cognition based on expectation. Narmour's theory draws on music theory, as well as ideas earlier proposed by Meyer (1956), who suggests that melodic cognition rests on innate perceptual principles of proximity, symmetry and similarity (cf. Schellenberg, 1996). Since Narmour makes concrete claims based on the score, the I-R model has subsequently been examined in detail as a model of melodic expectation (Huron, 2006). We summarise key aspects of the I-R model, as discussed in Schellenberg (1996).

The I-R model defines three processes, namely closure, implication and realisation. Closure entails the absence of expectation: in the context of two successive notes, closure is attained by factors including a shorter note followed by a longer note, the second note occurring on a stronger beat, or the second note conferring harmonic stability. Alternative factors which contribute to closure are a small interval succeeding a large interval, or changes in pitch direction.

The second process, implication, entails the presence of expectation and arises from the absence of closure. Based on three-note patterns, Narmour describes implication in terms of properties of registral direction (intervallic direction), intervallic difference (interval size), registral return (reverses in intervallic direction) and proximity (preference for small intervals). Based on exhausive enumeration of two-note patterns in terms of registral direction and intervallic differ-

ence, the I-R model defines expectations about the following note. The final process, realisation, describes the effect of fulfilled or violated expectation.

Early Probabilistic Approaches

With a view to characterising the predictability of music, Shannon's information theory (Shannon, 1948) provides a formalism for quantifying predictive uncertainty. As reported by Cohen (1962), information theory prompted early cross-fertilisation between music theory and psychology, specifically using information-theoretic measures computed on discrete-valued representations of music. Pertaining to music theory and psychology, Cohen identifies three investigation domains, namely 'analytic-synthetic', 'analytic', and 'synthetic'. The analytic-synthetic domain relates to the construction of statistical models using musical corpora for stylistically-informed music generation. Thus, we may view the analytic-synthetic approach as an expectation-modelling process for music composition. We may consider the remaining analytic and synthetic domains special cases of the analytic-synthetic domain: in synthetic investigations, the statistical model is presented in *ad hoc* fashion for music composition. Finally, in analytic investigations the statistical model is constructed on a corpus of music for musicological purposes. Common to many such approaches is the assumption of a Markov information source, whereby the conditional probability $P(s_i|s_{1:i-1})$ of the discrete-valued event s_i given the sequence of preceding events $s_{1:i-1} = (s_1, s_2, \ldots, s_{i-1})$ is equal to the conditional probability under finite context of length k,

$$P(s_i|s_{1:i-1}) = P(s_i|s_{i-k:i-1}) \tag{2.2}$$

where k is the order of the model. Conditional probabilities may be estimated using the empirical distribution of sub-strings of length $k+1$ (*n-grams*) in a corpus. For example, the respective conditional probabilities for first-order and second-order Markov models may be estimated using the empirical distribution of *bi-grams* and *tri-grams*.

In an analytic-synthetic investigation, Pinkerton (1956) obtains pitch bi-grams for a sample of 39 melodies. Using estimated pitch transition probabilities conditioned on metrical positions, Pinkerton obtains a simplified statistical model by considering the two most likely notes at each metrical position. Extending Pinkerton's method of determining first-order transition statistics, Brooks et al. (1957) obtain transition statistics up to the seventh order using a sample of 37 melodies, with further constraints restricting the set of generated melodies. In the work of Brooks et al. (1957), the use of n-grams to obtain transition statistics between notes and the assumption

of a Markov information source is analogous to the approach used in Shannon's estimate of information-theoretic entropy in natural language (Shannon, 1951).

We omit further discussion of early synthetic applications, since the employed model is typically strongly motivated (and thereafter supplemented) by the composer's aesthetic judgement (Cohen, 1962); we consider such topics beyond the scope of our discussion. Concerning analytic applications, Youngblood (1958) examines stylistic differences between Romantic era melodies and Gregorian chant, by computing first-order pitch transition statistics. In comparison to variations in entropies among individual Romantic era composers, Youngblood observes substantially lower entropies for Gregorian chant, with the requirement that both corpora are represented using a 12-tone scale. One may conclude that measures of musical expectation may be used to distinguish between musical styles, informal listening tests suggest however that higher-order statistics are necessary to generate stylistically convincing melodies (Cohen, 1962). As observed by Cohen, one may interpret entropy as quantifying the listener's average uncertainty when exposed to a musical stimulus, subject to the caveat that the listener has acquired the statistical structure of the entire stimulus. Related to Meyer's notion of the listener's expectations being revised as music unfolds in time (Meyer, 1956), Coons and Kraehenbuehl (1958) propose an 'information-dynamic' approach quantifying the listener's instantaneous predictive success, based on an evolving model. However, their approach is restricted to synthetic musical patterns and appears not to have been subsequently implemented.

Connectionist Approaches

As observed by Pearce (2007), the described probabilistic models and information-theoretic measures fell out of favour with the growth of artificial intelligence and cognitive science in the 1960s and the coinciding decline of behaviourism in psychology. Inspired by biological neural networks, artificial neural networks (ANNs) abstract the dynamics of biological neurons using non-linear functions, such that the parameters of such non-linear functions define individual units. Weighted input signals are supplied to units; outputs may be combined or directed to further units, as specified by the network's architecture. In a supervised learning setting, network parameters may be optimised using the backpropagation algorithm (Bryson et al., 1963). Two widely-applied classes of network architecture are feed-forward networks (FFNs) and recurrent neural networks (RNNs). In FFNs, input signals propagate to output units in an acyclic manner: input units are connected to hidden units; hidden units are in turn connected to output units. In

RNNs, the graph of unit connections contains cycles. For a more detailed discussion of ANN architectures, we refer to Bishop (2007).

Bharucha and Todd (1989) describe an ANN for predicting chords associated with scale degrees. Their approach is primarily based on a feed-forward architecture, whereby input and output units each correspond to major and minor chord functions. Given a chord sequence, successive chords are mapped to input unit activations; model weights are optimised by considering the mismatch between the immediately following chord and the network's output. The network models context of previously observed chords in the form of individually recurrent input nodes; weights of such connections are set to an identical constant within the range $(0, 1)$, thus implementing a decay of previous observations. Bharucha and Todd expose the network to artificial chord sequences generated using Piston's (1978) distribution of scale degree transitions. The authors observe that their network converges to a model of the conditional probability distribution associated with the chord sequences.

Mozer (1994) proposes an RNN for musical composition based on prediction. Analogous to Bharucha and Todd (1989), input units represent current musical events in a sequence; output units generate predicted events. Input and output units represent pitch and duration of notes, in addition to accompanying chords. In Mozer's model, hidden units are organised into multiple layers to enable a distributed representation of pitch, duration and harmonisation. A further layer of hidden units has recurrent connections as a representation of context as motivated by Bharucha and Todd (1989), however in Mozer's architecture recurrent connections are learnt, instead of being held constant. Based on artificial pitch sequences each on the scale of 10 elements, Mozer concludes that the architecture is able to learn musical structure at short time scales. Yet, Mozer observes limited success at predicting longer musical sequences, suggesting that long-range dependencies present a challenge for the model.

To address the challenge of learning long-range dependencies, alternative RNN architectures have been proposed. Eck and Schmidhuber (2002) apply the long short-term memory (LSTM) architecture (Hochreiter and Schmidhuber, 1997) for musical improvisation based on prediction; they consider melodic and harmonic accompaniment based on the 12-bar Blues form. The central concept in LSTM is the arrangement of units into memory cells capable of retaining an input signal indefinitely. For each cell, separate inputs pertain to storage, recall, and reset functionality. Eck and Schmidhuber observe that while long-range dependencies are learnt successfully, the

absence of variation in training data limits conclusions on the efficacy of their approach.

As reported by Bengio et al. (2013), RNN architectures offer a powerful framework for representing temporal dependencies, where specifically for the goal of learning long-range dependencies, model optimisation remains an ongoing research area. In a contrasting approach by Cherla et al. (2013), the use of a restricted Boltzmann machine (RBM) is proposed. Such a model may be viewed as an FFN where unit interactions represent probabilities. While the model as proposed does not represent long-range dependencies, it offers the advantage of efficient optimisation procedures (Hinton, 2002). Cherla et al. predict pitch sequences by supplying a fixed amount of context to input units, observing that the RBM competes with state-of-the-art n-gram approaches proposed by Pearce and Wiggins (2004). Bengio et al. (2013) evaluate a recurrent variant of RBM for melody prediction, observing encouraging results.

Recent Probabilistic Approaches

The work of Conklin and colleagues demarcates a renewed interest in information theory and probabilistic approaches for music prediction (Conklin and Cleary, 1988; Conklin and Witten, 1995). A limitation of earlier probabilistic approaches is that pitch and duration are either examined separately, or a single model of notes is constructed with pitch and duration jointly mapping to unique symbols. Such a cross-alphabet approach is preferable to the case of separate pitch and duration models, since it accounts for correlation between attributes. However, cross-alphabets may lead to inaccurate event probability estimates for objects with many attributes, since the cross-alphabet size grows exponentially with the number of considered attributes. A similar combinatorial issue arises for Markov models when increasing the context length: long contexts may lead to inaccurate estimates of conditional event probabilities, since the set of possible contexts grows exponentially with the context length. In both cases, the inaccuracy in probability estimates arises from sparse empirical distributions, where event probabilities of zero are underestimates.

Conklin and Cleary (1988) address the latter limitation by employing prediction by partial matching (PPM) (Cleary and Witten, 1984), a class of variable-order Markov model (VMM) (cf. Begleiter et al., 2004). In VMMs, predictions are formed using a set of Markov models with decreasing context lengths, where shorter context lengths are used when no observations exist for a given context. To address the first limitation relating to cross-alphabets, VMMs are estimated for separate 'viewpoints' such as pitch classes, intervals, and durations. To combine viewpoints

for prediction, the scheme proposed by Conklin and Witten (1995) involves weighting the probability distribution associated with each viewpoint according to information-theoretic entropy. In this manner, uncertain predictions are weighted less favourably than certain predictions. Finally, a set of viewpoint models is estimated separately for a corpus of musical pieces (the long-term model, LTM), as well as for the piece of music which is the subject of prediction (the short-term model, STM). Expanding on the described LTM/STM approach, Pearce and Wiggins (2004) investigate influences of using VMM variants, as well as methods for estimating and combining the LTM and STM. Pearce and Wiggins consider the problem of predicting pitch sequences based on a single viewpoint representation, where they improve on multiple-viewpoint pitch prediction accuracies reported by Conklin and Witten (1995). The LTM/STM approach with multiple viewpoints is subsequently proposed as a model of listeners' melodic expectations (Pearce and Wiggins, 2006; Pearce et al., 2010).

Concerning the prediction of musical harmony, Ponsford et al. (1999) use tri-gram statistics to estimate chord transition probabilities on a musical corpus; estimated probabilities are then used to generate chord progressions. Ponsford et al. base their approach on a key-invariant representation, where chords are represented as lists of scale degrees; major and minor modes are modelled separately. Based on VMMs and for the purpose of generating harmonisations with respect to a melody, Whorley et al. (2013) propose a multiple viewpoint system based on VMM modelling. Whorley et al. propose methods for reducing computational complexity in the harmonisation problem, including incorporating domain knowledge on melodic ranges.

Assayag et al. (1999) propose an approach for automated musical improvisation based on cross-alphabets. For the described application, the limitations of cross-alphabets appear to be of secondary concern, since the improvisation is based on music with few melodic voices; moreover attaining high predictive accuracy is subsidiary to the goal of novelty in improvisations, which is evaluated subjectively. Assayag et al. form predictions using an adaptation of the widespread Lempel-Ziv (LZ) string compressor (Ziv and Lempel, 1977), which may be considered a class of VMM (cf. Begleiter et al., 2004). In contrast to PPM which specifies an upper bound on VMM order, LZ specifies no such bound (outside implementation constraints). Thus, Assayag et al. motivate their choice of LZ compression to better account for long contexts when forming predictions.

Information-Theoretic Measures of Predictability

The hitherto-described approaches focus on the use of predictive models: where information theory is used to quantify predictive uncertainty, investigations are confined to the measures originally proposed by Shannon (1948) for discrete noiseless systems. Abdallah and Plumbley (2009) compare related information-theoretic measures of predictability in sequences, as apprehended by an evolving probabilistic model. By grouping a sequence of events into a perceptual past, present and future, the authors propose and contrast measures of instantaneous predictive success; such measures are based on the notion that the listener forms predictions using the perceptual past and revises the internal model, as events unfold in time. In particular, the authors propose *predictive information*, which pertains to the amount of information the present conveys about the future, given the past. The authors propose average predictive information (the *predictive information rate*, PIR) as a measure of sequential complexity: the PIR attains maximal value for sequences which have intermediate statistical regularity on the continuum of uncorrelated and deterministic sequences. Thus the authors propose PIR as a measure of cognitive interestingness, considering that for sequences with high PIR, observations in the perceptual present are relatively informative about the perceptual future. Using a first-order Markov model and based on an analysis of minimalist compositions using monophonic pitch sequences, Abdallah and Plumbley qualitatively examine instantaneous measures of information for any correspondence with temporal structure in the compositions. As observed, all measures appear to show some amount of correlation with expert-annotated structural boundaries; notably peaks in the *log-loss* (cf. Chapter 3) and predictive information measures coincide with sectional boundaries. Abdallah and Plumbley note that abstraction and generality are attributes of their information-theoretic approach: the examined quantities are not restricted to perceptual entities such as notes. In addition, the authors note that the measures are not restricted to the first-order Markov model considered in the investigation.

Continuous-Valued Approaches

The preceding models of expectation are strictly applied to score representations of music, where both pitch and duration are discrete-valued. We refer to such representations as *symbolic* representations; where note durations and onsets are initially continuous-valued, they are subsequently quantised to obtain discrete-valued representations (Assayag et al., 1999). Pertaining to musical audio, Large (2000) proposes a model of metre as a continuous-time system of oscillators. In the

presence of an input signal, oscillators may entrain to a rhythm, resulting in oscillations which are phase-locked with rhythmic onsets, such that oscillations persist after initial entrainment. Oscillations are proposed to embody the perception of metre, implying predictions about successive rhythmic events; we may thus consider Large's approach a model of expectation in time. Cont (2008, p. 121) incorporates an oscillatory model into an audio-to-score alignment system, where oscillations are used to form anticipatory actions and assist in inferring the current score position.

Hazan (2010, p. 61) proposes a model of expectation for musical audio, which addresses the problem of learning classes of acoustic events and their rhythmic onsets. Hazan's model is based on discrete-valued representations of event onset intervals, sequences of which are learnt online using a VMM, similar to the STM approach described by Conklin and Witten (1995). Inter-onset intervals are quantised using an online variant of the K-means clustering algorithm. As a representation of musical timbre in acoustic events, Mel-frequency cepstral coefficients (MFCCs, cf. Rabiner and Juang (1993)) are employed, which are assigned to clusters analogously, using online K-means. The number of clusters is optimised using goodness-of-fit criteria in an initial learning phase, where the system is exposed passively to a musical signal. With a view to forming predictions, Hazan considers three possible strategies for modelling the resulting sequences of quantised inter-onset intervals and timbral clusters. These include independent models, cross-alphabets, and separate inter-onset models conditioned on timbral clusters. Hazan performs an evaluation of prediction behaviour, examining the influence of exposure to percussive and melodic musical signals. Exploratory analysis of instantaneous entropy suggests that the model successfully learns temporal structure in the considered signals, and that predictive uncertainty may be used to infer repetition structure. Empirical analysis demonstrates that the system successfully predicts percussive events. Analogous approaches by Marchini and Purwins (2010); Marxer and Purwins (2010) explore the use of online hierarchical clustering techniques. Contrasting with the previously described probabilistic approaches, Dubnov et al. (2007) identify repetition structure in audio by combining feature quantisation with methods for constructing finite-state automata (Allauzen et al., 1999). Dubnov et al. perform concatenative sound synthesis by randomly traversing learnt automata. Finally, Stark and Plumbley (2012) propose an approach for real-time musical accompaniment based on sequence alignment of audio features. This approach requires no quantisation, relying instead on determining similarities between beat-synchronous feature vectors in Euclidean space.

Information-Theoretic Approaches at the Sub-symbolic Level

The information-theoretic measures proposed and evaluated by Conklin and Witten (1995); Pearce and Wiggins (2004); Abdallah and Plumbley (2009) are based on symbolic representations of music; conceptually, the continuous-valued approaches proposed by Hazan (2010); Marchini and Purwins (2010); Marxer and Purwins (2010) differ only in the additional use of quantisation to obtain symbolic representations from audio. Adopting a sub-symbolic approach (i.e. operating above the musical surface, thus relating to perceptual processes rather than cognitive processes; cf. Section 2.2.3), Dubnov (2006) proposes 'anticipation profiles' which quantify instantaneous, short-term predictability in audio. Rather than perform transcription and quantify predictability in symbol sequences, Dubnov attempts to quantify the type of predictability which distinguishes musical signals at short time scales both from noise and straightforward determinism. To this end, Dubnov's *information rate* (IR) is the reduction in uncertainty about the present obtained when accounting for the past[1]. The IR is minimal for signals which are either random or deterministic. Dubnov shows that for stationary Gaussian processes such as white noise autoregressive processes (cf. Lütkepohl, 2005, p. 4), the IR may be computed as the *spectral flatness* from the signal's power spectrum. To deal with complex audio mixtures, Dubnov sums IR across decorrelated spectral features. Thus computed, Dubnov performs an exploratory analysis of IR for a natural sound-scape, versus synthetic noise. Computing IR over the entire duration of the signal, results suggest that IR successfully quantifies temporal regularity. In a subsequent exploratory analysis involving musical audio, Dubnov obtains instantaneous power spectra by applying a 3 second sliding window to spectral features. Plots of instantaneous IR suggest that the measure captures information on temporal structure distinct from alternative statistics such as signal energy. In a study involving contemporary orchestral music, Dubnov et al. (2006) observe that IR significantly correlates with real-time measures of affective response in listeners. Moreover, a combination of IR and signal energy yields substantially higher correlation with affective response, compared to signal energy alone.

The proposed formulation of IR assumes that signals are stationary, which Dubnov argues is reasonable for musical signals which are assumed to be locally stationary within the specified 3 second window. Notably, predictability is quantified only with respect to short time scales. Dubnov (2008) subsequently estimates instantaneous predictability while modelling non-stationarity

[1] Dubnov's IR differs from Abdallah and Plumbley's PIR (cf. Abdallah and Plumbley, 2009).

between successive window positions. To this end, Dubnov decomposes IR into a local *data-IR* term (as previously defined), in addition to a global *model-IR* term. The model-IR is estimated by applying quantisation to blocks of features represented by their average, assuming a first-order Markov model for transitions between quantised features. Exploratory analysis using piano recordings and using a 5 second window suggests that inspection of separately visualised data-IR and model-IR may be used to identify interest points in a piece of music: data-IR may be used to identify repetitive versus irregular melodic patterns, whereas model-IR may be used to identify novelty versus repetition structure at longer time scales. The analysis of combined data-IR and model-IR is suggested for future work.

Recently, Abdallah and Plumbley (2013) have extended the discrete-valued treatment of PIR (Abdallah and Plumbley, 2009) to the continuous-valued case. For PIR, the authors derive expressions for Gaussian processes, involving power spectra. For the particular case of white noise autoregressive processes, Abdallah and Plumbley derive straightforward closed-form expressions for the PIR, involving autoregressive coefficients. Abdallah and Plumbley then propose a Bayesian method for estimating autoregressive parameters in online fashion. Assuming local stationarity, the approach relies on maintaining a record of past observations with exponential memory decay. Adopting Dubnov's approach of decorrelating spectral features, Abdallah and Plumbley apply their Bayesian approach to a minimalist percussion piece. Based on the evolving model and maintaining around 12 seconds of past observations, the authors then compute instantaneous measures of predictability, including PIR and IR. Exploratory analysis of individual information measures suggests that multiple measures convey the presence of structural boundaries in the piece of music, mirroring the results reported by Abdallah and Plumbley (2009) for symbolic representations.

2.3.4 Summary and Conclusions on Musical Expectation

This far, we have discussed the role of expectation in musical pitch and rhythm. We observe that for both phenomena, expectation plays an important role in the cognition of musical structure. In particular, cognitive representations of pitch and rhythm involve statistical learning (Huron, 2006). In our discussion of expectation models, we identify two broad categories of approach, based on statistical and connectionist models, respectively. A successful statistical approach to quantifying predictive uncertainty involves computing information-theoretic quantities and assuming a Markov model (Conklin and Witten, 1995; Pearce and Wiggins, 2004). Whereas

early methods primarily involve mono-alphabetic representations, more recent methods employ complex alphabets when dealing with symbolic (i.e. discrete-valued, score-level) representations of music, or use quantisation when dealing with continuous-valued features.

We seek to establish whether an approach inspired by expectation modelling might be useful in determining similarity in musical audio. This aim is motivated by the observation that expectation is important in determining musical structure (Wiggins, 2007). Our working hypothesis is that measures of predictability might be used to identify music with similar temporal structure, where we extend our hypothesis to include sub-symbolic structure. In this view, we deem sequences musically similar, if they incur similar amounts of predictability. As the conceptual framework in our enquiry, following the described statistical approaches we consider an information-theoretic approach and evaluate measures of predictability in musical audio.

Concerning our approach, a caveat relates to the observation that the reviewed models of musical expectation are primarily based on symbolic representations of music. In fact, the domain in which the information-theoretic approach has been validated—in behavioural and neurophysiological terms—is melody (Pearce and Wiggins, 2006; Pearce et al., 2010); it remains to investigate the psychological plausibility of this approach for alternative musical phenomena. Since we consider measures of predictability as a conceptual framework, psychological plausibility is not of primary concern in our investigations. From an engineering perspective, we note that one possible approach to computing information-theoretic measures involves combining quantisation with a discrete-valued statistical model (Hazan, 2010). An alternative approach involves prediction in the continuous domain, as proposed by Abdallah and Plumbley (2013).

A limitation of existing audio-based approaches is that evaluations involving information-theoretic measures are primarily confined to exploratory analysis (Dubnov, 2006; Hazan, 2010; Abdallah and Plumbley, 2013). As we ascertain in Section 2.4, few investigations have applied such approaches in music content analysis. Considering our aim of evaluating information-theoretic measures of predictability, this thesis seeks to redress the gap in the existing literature.

Concerning the use of information-theoretic measures of predictability, we observe that such an approach offers the advantage of *abstraction* and *generality* (Abdallah and Plumbley, 2009). Abstraction refers to the notion that the meaning of observations is interchangeable, thus the approach relates to patterns of observations, rather than the observations themselves. Relating to our work, abstraction allows us to analyse patterns in both quantised and non-quantised

continuous-valued sequences, using the same conceptual framework. Secondly, generality refers to the notion that probabilistic models may be interchanged, while retaining the interpretation of examined measures. We consider *expressiveness* a third advantage of the approach, since it admits a multitude of measures which may potentially be used to quantify predictability.

Concerning potential tasks in music content analysis, we conjecture that an audio-based approach close to the musical surface might prove successful, providing we can obtain a relevant discrete-valued representation from the audio signal. One task which operates close to the musical surface is version identification (cf. Serrà, 2011), where chroma features (Fujishima, 1999; Bartsch and Wakefield, 2001) have been frequently used as a representation of harmonic content in music, and where similarity is typically determined by comparing between audio tracks sequences of chroma features. For version identification, an information-theoretic approach offers the potential for contrasting discrete-valued and continuous-valued measures of pairwise predictability hitherto not considered, and for evaluating their behaviour as similarity measures. With a view to such an evaluation in Chapter 4, we review work on version identification in Section 2.4.2.

Concerning sub-symbolic measures of predictability, we identify a potential use in low-specificity similarity tasks such as genre classification (cf. Fu et al., 2011b), which use summary statistics computed in track-wise manner from sequences of audio features. Since predictability measures account for temporal structure, they might serve to complement alternative summary statistics where temporal structure is discarded. The track-wise approach differs from the pairwise approach suggested above, in that we consider the obtained measures of predictability our elements of comparison, thus we compare tracks in terms of their individual measures of predictability. We note that the track-wise approach is not restricted to the sub-symbolic level, thus we might attempt to quantify predictability close to the musical surface. On the other hand, we note that existing statistics for low-specificity similarity attempt to capture information at the sub-symbolic level (Fu et al., 2011b), suggesting that a sub-symbolic approach is more likely to be successful. We review work on low-specificity similarity tasks in Section 2.4.3, before providing an account of our work in Chapter 5.

2.4 Music Similarity Computing

2.4.1 Audio Features

In music content analysis, methods for characterising musical similarity are principally based on obtaining sequences of features from audio, with such features obtained from time-frequency representations. As discussed by Fu et al. (2011b), one possible feature taxonomy distinguishes between low-level and mid-level features. Low-level features are obtained by applying spectral analysis to frames on the scale of 10ms–100ms, thereafter statistics may be computed which characterise the magnitude spectrum at the same time scale. Low-level features have been used as a representation of musical timbre, above the musical surface. Mid-level features in contrast are based on longer time scales, and aim to represent phenomena close to the musical surface, such as rhythm and harmony. Meanwhile, mid-level features may be considered an alternative to symbolic features, the latter which are more challenging to obtain (Marolt, 2008).

Tzanetakis and Cook (2002) have proposed the use of MFCCs as a representation of musical timbre for genre classification. In MFCCs, log-transformed magnitude spectra obtained using the discrete short-time Fourier transform (STFT) are divided into sub-bands and weighted. The spacing of sub-bands is linear at low frequencies, and logarithmic at high frequencies, based on a perceptually motivated scale. After dividing and weighting magnitude spectra, the resulting signal is cosine-transformed to obtain MFCCs. For speech and music, we may consider the cosine-transformation a decorrelation step which approximates principal components analysis (PCA), thus the resulting coefficients in ascending order explain a decreasing amount of variance in the log-transformed magnitude spectrum (cf. Logan, 2000). The use of 13 MFCCs is typical in music content analysis (Lartillot and Toiviainen, 2007).

Alternative low-level features are based on computing scalar-valued statistics on frame-wise estimates of magnitude spectra (Tzanetakis and Cook, 2002). For example, the *spectral centroid* may be defined as a weighted average of the magnitude spectrum, the *spectral flux* may be defined as the sum of squared errors between successive magnitude spectra, and the *spectral rolloff* may be defined as a specified percentile of the magnitude spectrum. Finally, it should be noted that spectral analysis is not confined to approaches based on the STFT, for example Li and Ogihara (2006) propose features based on the discrete wavelet transform.

A mid-level feature, chroma features (Fujishima, 1999; Bartsch and Wakefield, 2001) have been proposed as a representation of harmonic content. Chroma features quantify energy dis-

tributions across octave-folded bands, using the 12 pitch classes in the chromatic scale to map frequency bands to discrete chroma bins. A variety of approaches for computing chroma have been proposed: Fujishima (1999) computes an STFT and for each frame sums spectral energies in octave-folded frequency bands, to obtain 12-component features. Bartsch and Wakefield (2001) propose a similar approach, but instead of using a constant frame-rate the authors apply beat tracking (Dixon, 2000) and obtain beat-synchronous features. Both approaches assume an equal-tempered scale with respect to a fixed pitch reference. In contrast, Gómez and Herrera (2006); Ellis and Poliner (2007) account for deviations from the pitch reference, by initially using 36 octave-folded frequency bands, and by heuristically shifting frequency bands, respectively.

In addition to computing beat-synchronous features, there exist mid-level features which attempt to describe rhythmic content exclusively. Given an audio signal, Tzanetakis and Cook (2002) detect periodicities by selecting peaks in the sample autocorrelation of the audio signal envelope. Another approach consists in determining individual musical onsets and clustering obtained onset intervals (Dixon, 2001). Such methods may be used to obtain global or instantaneous tempo estimates, by using further heuristics (cf. Müller et al., 2011).

2.4.2 Version Identification

Version identification systems commonly assume that sequential pitch content is preserved among versions of the same piece of music (Serrà, 2011). Owing to this assumption, version identification relies on predominant melody extraction to represent melodic content, or extraction of chroma features to represent harmonic content. Shown in Figure 2.2, feature extraction may be followed by additional processing to obtain a summary feature representation. Additional processing stages aim to adjust for any variation in musical key or tempo between versions. Pairwise sequence matching is then applied, which results in a measure of pairwise similarity between tracks. We discuss each of the stages separately, in similar spirit to Serrà (2011). Our discussion differs in that we focus on methods for pairwise sequence matching and computing pairwise similarity, the subject of our own work.

Note from Figure 2.2 that obtained pairwise similarities are typically used to quantify retrieval performance, using a suitable statistic. Thus we may view Figure 2.2 as a variant of Figure 1.1 omitting model estimation and including additional steps, in place of straightforward feature extraction and obtaining a track-wise representation.

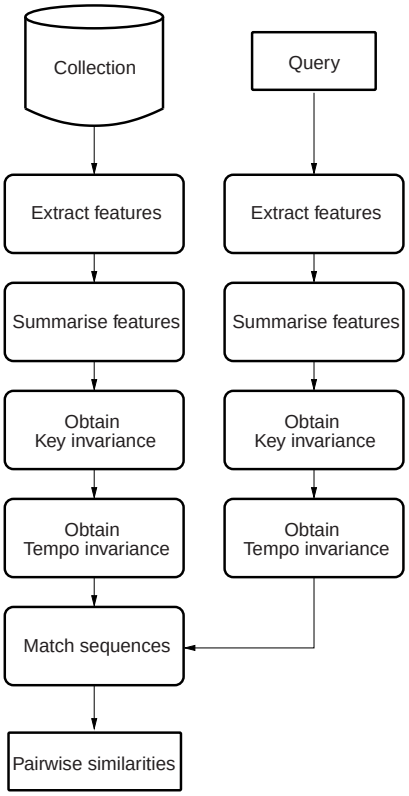


Figure 2.2: Stages involved in a typical version identification system.

Feature Extraction and Summarisation

In an approach based on predominant melody extraction, Tsai et al. (2005) assume that solo vocal content in popular music contains melodic material relevant to identifying cover versions of a song. Following this assumption, the authors segment audio into vocal and non-vocal regions, using a classifier based on MFCCs. Non-vocal regions longer than two seconds are then discarded. To perform predominant melody extraction, Tsai et al. apply the STFT and perform spectral peak picking, where the approach sums energies across harmonics to improve the robustness of fundamental frequency estimation. Successively repeating pitch candidates are discarded, since such notes are likely to have been produced by instrumental accompaniment. Additional heuristics constrain the range of admissible durations and pitches.

Marolt (2008) describes a mid-level approach where features encode multiple salient melodic lines, rather than the predominant melody. The approach derives from Klapuri (2006), whereby acoustic modelling is first applied to remove noise content in the signal. In a two-stage process, a set of pitch candidates is then formed using spectral peak picking, which results in a set of melodic lines where each pitch has an associated salience value. To obtain the mid-level

representation, melodic lines are then pruned by applying a threshold to average salience values.

Since reliable melody extraction presents a challenge in complex audio mixtures, most recent version identification systems rely on mid-level representations, notably chroma-based representations of harmonic content. We may distinguish between approaches which retain the continuous-valued representation of chroma for pairwise similarities, and those which use quantised chroma features. Concerning quantisation-based methods, Casey and Slaney (2006) propose an approach based on the K-means clustering algorithm. Here, the codebook is estimated using training data, with codebook sizes in the range $[8..64]$. An alternative, unsupervised approach consists of applying a threshold to each chroma vector, which results in a set of 2^{12} possible symbols for 12-dimensional chroma features (Tabus et al., 2012). A caveat against using the latter approach is that the codebook is comparatively large and may contain redundant codewords; an alternative approach uses musical knowledge to define a set of 793 chroma codewords (Kurth and Müller, 2008). We may consider chord recognition a particular case of quantisation, where the set of symbols is relatively small and musically motivated. Lee (2006) performs chord recognition using a 36-state hidden Markov model (HMM), trained on labelled chord sequences. States are defined for three possible chord types (major, minor, diminished), for each of 12 possible root notes. Bello (2007) and Ahonen (2009) propose a similar approach using 24-state and 12-state HMMs, respectively based on two chord types (major, minor) and a single chord type ('indeterminate mode'). An alternative approach relies on matching chroma vectors with binary templates (Martin et al., 2012; Khadkevich and Omologo, 2013).

Tempo Invariance

To deal with tempo variation between renditions of a piece of music, beat-synchronous features have been widely applied (Bello, 2007; Ellis and Poliner, 2007; Serrà et al., 2008; Bello, 2011; Bertin-Mahieux and Ellis, 2011, 2012; Khadkevich and Omologo, 2013). On the one hand, this approach presents a potential caveat: Bello (2007); Serrà et al. (2008) attribute reduced performance to unreliable beat tracking. On the other hand, the approach potentially yields features which are insensitive to both local and global tempo changes. In addition, beat-synchronous representations may yield shorter sequences compared to frame-based representations, thus reducing the computational cost of pairwise sequence matching. An alternative approach consists of re-sampling the chroma sequence to a lower frame-rate (Serrà et al., 2008; Bello, 2011); however in contrast to beat-synchronous features this approach is unable to adjust for local tempo

variation. A further method which aims at tempo-insensitivity involves interpreting the chroma sequence as a two-dimensional signal: Jensen et al. (2008) compute the two-dimensional sample autocorrelation; tempo-insensitivity is subsequently obtained by computing a weighted sum of autocorrelations across time lags, with weights defined by exponentially distributed bands.

Key Invariance

The concept of musical key is relevant in cover song identification, since cover versions of a song may involve chromatic transposition, i.e. the shifting of pitch by a specified interval. Since pitch class is a modulo 12 representation of chromatic pitch, transposing a set of pitches implies applying a circular shift to associated pitch classes. Given two chroma feature sequences, one approach for handling transposition relies on repeated pairwise sequence matching, by considering all 12 possible circular shifts of chroma feature components and minimising the similarity measure (Ellis and Poliner, 2007; Kurth and Müller, 2008; Marolt, 2008). With a view to reducing computational cost, the required number of circular shifts may be determined using summarised chroma sequences: Serrà et al. (2008) proposes the optimal transposition index (OTI), based on averaging both chroma sequences across time, before computing the inner product between the resulting chroma vectors. The inner product is then optimised with respect to rotations of the query features. An alternative approach relies on key estimation, as proposed by Gómez and Herrera (2006); this approach is however observed to be less robust compared to the OTI (Serrà et al., 2008). Bello (2007) proposes a similar approach to the OTI, using normalised chord histograms. Finally, (Marolt, 2008; Bertin-Mahieux and Ellis, 2012) compute two-dimensional power spectra, which have the property of shift-invariance. Applied to chroma features, the resulting power spectra are therefore key-invariant.

Pairwise Sequence Matching

A variety of version identification approaches are based on computing pairwise alignments between continuous-valued chroma sequences. Following Foote's (2000) method of applying dynamic time warping (DTW, cf. Sakoe and Chiba, 1978) to spectral energy features, Gómez and Herrera (2006) apply DTW to chroma features. Given two feature sequences, DTW is based on computing a similarity matrix between feature sequences, where each entry records the similarity between two feature vectors respectively from each sequence, using a specified distance function such as the Euclidean distance. The similarity matrix is used to compute an optimal, cumulative similarity score between both sequences, by specifying additional costs for inserting or deleting

feature vectors from one of the sequences. The associated set of insertions and deletions defines a global alignment between the two sequences. Computed using frame-based feature representations, DTW is thus able to account for tempo variation between musical renditions by inserting or deleting feature vectors, without resorting to beat-synchronous features. Serrà et al. (2008) evaluate the performance of alignment techniques based on DTW, which place additional global or local constraints on admissible insertions and deletions. Whereas global constraints relate to the total number of permitted insertions and deletions, local constraints relate to the amount of permitted time dilation and compression. A further notable difference to Foote (2000); Gómez and Herrera (2006) is the use of binarised similarity matrices, obtained by thresholding pairwise distances between feature vectors. In addition, the authors propose a local alignment technique based on the Smith-Waterman algorithm (Smith and Waterman, 1981), which incorporates local alignment constraints. Given a query sequence and a comparison sequence, local alignment determines the optimal set of insertions and deletions from the query, with respect to all possible contiguous sub-sequences of the query and comparison sequence. In this manner, local alignment is able to account for structural changes between musical renditions, such as omitted verses in cover songs.

Subsequent investigations by Serrà et al. (2009) extend the notion of similarity matrices used for sequence alignment in the preceding investigations, in that time-lagged chroma vectors are combined to form higher-dimensional temporal features, using the process of *time delay embedding* (Takens, 1981). In this way, each resulting vector captures information on temporal structure, analogous to n-grams formed from a sequence of symbols (cf. Section 2.3.3). By thresholding pairwise distances between time delay embedded vectors, Serrà et al. (2009) obtain *recurrence plots*. Similarity between query and comparison chroma sequences is subsequently quantified using statistics of diagonal paths in the associated recurrence plots. Local or global tempo variation is accounted for, by allowing paths to vary in curvature or angle, respectively.

In an alternative approach, Serrà et al. (2012) propose to use a measure of pairwise predictability between frame-based feature sequences as a means of quantifying similarity. The authors apply non-linear prediction techniques to time delay embedded features. Given a query and a comparison sequence, the approach relies on estimating a predictive model with respect to the comparison sequence. The estimated model is then used to obtain a sequence of predictions about successive elements in the query sequence; this sequence of predictions is thereafter com-

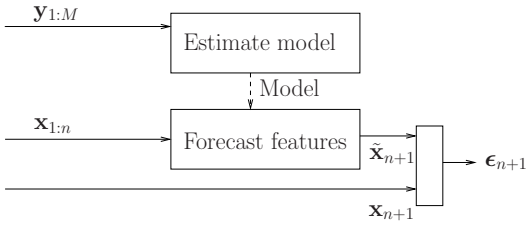


Figure 2.3: Cross-prediction. Sequence $\mathbf{y}_{1:M} = (\mathbf{y}_1, \ldots, \mathbf{y}_M)$ serves as model input, the observation context $\mathbf{x}_{1:n}$ thereafter forms basis of prediction $\tilde{\mathbf{x}}_{n+1}$. Quantity $\boldsymbol{\varepsilon}_{n+1}$ denotes the prediction error.

pared to actual elements in the query sequence, using an error statistic. Figure 2.3 illustrates the described approach schematically, which Serrà et al. term *cross-prediction*. The use of cross-prediction for version identification is based on the intuition that a similar comparison sequence is a feature sequence whose model facilitates accurate prediction of the query sequence.

Using signal processing techniques, Ellis and Poliner (2007) apply two-dimensional cross-correlation as a measure of similarity between beat-synchronous chroma sequences. Specifically, the authors determine similarity by applying peak-picking to sample cross-correlations normalised with respect to the shorter of the two sequences; in this manner obtained correlation values fall within the unit interval. A caveat exists in that strong correlations may arise from a single chroma component across a large range of lags, where in contrast matching chroma sequences typically yield strong correlations across multiple chroma components, for specific lags. Owing to the latter observation, the authors seek to identify maxima with sharp peaks by high-pass filtering the sample cross-correlation. A further approach using signal processing techniques is proposed by Jensen et al. (2008). As previously described, the approach uses two-dimensional sample autocorrelations of chroma sequences to obtain tempo-invariant representations. Similarly, (Bertin-Mahieux and Ellis, 2012) compute two-dimensional power spectra to obtain key-invariant chroma representations. In both cases, transformed feature sequences are subsequently compared using the squared Euclidean distance.

The hitherto-described pairwise sequence matching techniques are applied to continuous-valued features. Based on a discrete-valued feature representation, a number of approaches apply sequence alignment techniques analogous to those described previously: Tsai et al. (2005); Lee (2006); Bello (2007) perform global alignment; for the considered discrete-valued approaches, the distance function previously described for DTW instead specifies the cost of symbol substitution, instead of the distance between feature vectors. Casey and Slaney (2006); Khadkevich and

Omologo (2013) apply the string edit distance, which we may consider a particular case of DTW, where symbol substitutions incur unit cost. Martin et al. (2012) instead perform local alignment, using a computationally efficient heuristic approach (Altschul et al., 1990).

Another discrete-valued approach relies on the normalised compression distance (NCD) (Li et al., 2004) to quantify similarity. We discuss the NCD in detail, in chapters 3 and 4. Given two strings x, y, the NCD is defined as

$$\mathrm{NCD}(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \tag{2.3}$$

where $C(\cdot)$ denotes the number of bits required to encode a given string, using a specified string compressor. Moreover, $C(xy)$ denotes the number of bits required to encode the concatenation of strings x, y. The NCD is based on the intuition that for similar strings of equal length, the value of $C(xy)$ approximates both $C(x), C(y)$, whereas for dissimilar strings of equal length, we have $C(xy) \gg \max\{C(x), C(y)\}$. For typical compression algorithms, Equation 2.3 yields values approximately within the unit interval (Li et al., 2004).

Ahonen computes the NCD between quantised feature sequences for version identification (Ahonen, 2009, 2010, 2012). The approach draws on earlier work which applies the NCD to symbolic representations of music (Li and Sleep, 2004; Cilibrasi and Vitányi, 2005; Ahonen et al., 2011), and to quantised audio features (Li and Sleep, 2005) for genre classification. Using chord identification to quantise chroma features, Ahonen (2009) evaluates the effect of interchanging compressors on version identification accuracy; the comparison includes PPM (Cleary and Witten, 1984), Burrows-Wheeler (BW) compression (Burrows and Wheeler, 1994) and LZ compression (Ziv and Lempel, 1977). Ahonen (2010) proposes to average NCDs computed using multiple discrete-valued representations, with PPM used for string compression. Ahonen (2012) proposes chroma-derived representations which are then quantised and compressed using BW compression. Tabus et al. (2012) propose a similar approach to Ahonen based on quantising chroma-derived representations, using an alternative compression-based similarity measure to the NCD.

In a contrasting approach, Bello (2011) computes recurrence plots obtained for individual chroma sequences in track-wise manner, rather than pairs of chroma sequences as proposed by Serrà et al. (2009). Such *self-recurrence plots* are then compared using the NCD, which Bello proposes as a measure of structural similarity between two pieces of music. Finally, Silva et al. (2013) propose related measures of structural similarity by applying video compression to recur-

rence plots, using an alternative compression-based measure to the NCD.

Large-Scale Version Identification

A number of recent investigations are concerned with the potential use of version identification for large-scale music collections containing millions of tracks. For such collections, it may be infeasible to perform computationally expensive comparisons involving every track in the collection, given a query. An example for a comparatively expensive algorithm is alignment based on DTW, which for a single pairwise comparison involving feature sequences with respective lengths N, M has $O(NM)$ time complexity[2].

Kurth and Müller (2008) propose an approach based on an inverted file index (cf. Clausen and Kurth, 2004). Given a codebook and a collection of quantised feature sequences, this data structure maps each codeword to the set of sequences in which the codeword occurs. Given a query sequence, the inverted file index is used to map query codewords to sets of candidate sequences. Retrieval is subsequently performed by intersecting sets of candidate sequences, an operation which may be performed computationally efficiently using sorted lists (Clausen and Kurth, 2004). Kurth and Müller subsequently propose a fuzzy matching scheme, which permits the retrieval of candidate sequences with mismatching symbols. The authors examine a scenario in which an indexing stage is performed offline, prior to a query and retrieval stage, which is performed online. The authors propose a *filter-and-refine* approach (cf. Schnitzer et al., 2009), whereby a more computationally expensive matching is subsequently applied to the set of retrieved items in the collection.

Alternative large-scale approaches relate to locality-sensitive hashing (LSH), (cf. Slaney and Casey, 2008). The central concept in LSH is a projection operator which reduces the dimensionality of a feature space. Given two feature vectors and a similarity measure, a suitable projection operator maps the feature vectors to lower-dimensional vectors, such that the similarity between low-dimensional vectors approximates the similarity between the original feature vectors. Next to the projection operator, a hash function maps the low-dimensional subspace to a set of hash values. Thus, LSH aims to ensure that two similar feature vectors map to the same hash value with high probability. The hash function allows a set of feature vectors to be stored in a lookup table indexed by hash values. Similar entries in the lookup table are subsequently identified by obtaining the hash value for a query feature vector, an operation which may be performed in $O(1)$

[2]Whereas DTW as originally proposed by Sakoe and Chiba (1978) is relatively computationally expensive, we note that there exist computationally efficient variants, cf. Rakthanmanon et al. (2012).

time (Slaney and Casey, 2008).

Casey et al. (2008a) propose version identification as nearest-neighbour retrieval using a similarity threshold. Given a query and a comparison track, two sets of fixed-length chroma fragments are obtained by windowing track-wise chroma sequences. Pairwise similarity is then quantified as the number of fragments in the query track for which there exists a proximate fragment in the comparison track. The authors define proximity in terms of Euclidean distance falling below a specified threshold. Importantly, the approach of using the Euclidean distance allows LSH to be employed, as subsequently investigated by Rhodes et al. (2010). Specifically, Rhodes et al. consider the approach of random projections, in which a set of inner products is computed between a feature vector and points drawn from a multivariate Gaussian distribution. The number of computed inner products determines the dimensionality of the feature subspace after projection; hashing is subsequently achieved by quantising the subspace. Marolt (2008) considers a similar LSH approach for estimating cosine distances; this approach relies on computing the sign of inner products between feature vectors and random points. In both cases, the feature vectors in question are vectorised fragments of audio feature sequences.

More recently, Khadkevich and Omologo (2013) propose a filter-and-refine approach for version identification, based on LSH. Here, each track is represented as a distribution of chords, with chord distributions compared using the L1 distance. The authors use LSH to obtain a set of nearest neighbours with respect to a query; this set of nearest neighbours is subsequently ranked using the edit distance. In an alternative hashing approach, Bertin-Mahieux and Ellis (2011) binarise and window chroma sequences using an adaptive thresholding technique. Each fragment is subsequently mapped to a hash value, based on an encoding of chroma peak locations. Finally, Martin et al. (2012) match chord sequences using heuristic local alignment (Altschul et al., 1990). Here, in an initial step, n-grams obtained for a query sequence are compared to n-grams obtained for all sequences in a collection. By locating matching n-grams in the collection, the algorithm identifies short sequences with exact matches, which form seeding points for subsequent local alignment. In this manner, the set of sequences considered for local alignment is constrained.

Discussion

In the preceding, we have reviewed existing approaches to version identification. With particular attention to methods for pairwise sequence matching, we may distinguish between methods based on strings, which contrast with methods based on sequences of continuous-valued audio features.

We note a number of discrete-valued approaches which are based on the NCD. On the one hand, this approach is straightforward to compute: subject to the compression algorithm employed, the approach may be considered parameter-free (Li et al., 2004; Sculley and Brodley, 2006). It thus contrasts with DTW, which requires a cost function to be defined, in addition to a distance measure. In addition, DTW has time complexity $O(NM)$ with respect to sequence lengths N,M. In contrast, string compression may be performed in linear time with respect to sequence lengths, thus following Equation 2.3 pairwise string comparison has time complexity $O(N+M)$.

We note that the NCD may be considered an information-theoretic measure, if we consider the associated string compression from an information-theoretic perspective (cf. Begleiter et al., 2004). As we review in Chapter 3, NCD is motivated as an approximation of an uncomputable, optimal similarity measure (Li et al., 2004), which we may consider a measure of pairwise predictability between sequences. Although we may conceive of related measures of similarity (Tabus et al., 2012; Silva et al., 2013), to date no detailed comparison of related methods has been performed. As outlined in Section 2.3.4, we use our information-theoretic framework to perform such an investigation.

With regard to discrete-valued approaches, Serrà (2011, p. 28) observes that while quantisation yields efficient representations of pitch, its effect on version retrieval accuracy remains to be explored in detail. Particular to the NCD, we note that existing evaluations are predominantly based on datasets of the scale of 10^2 tracks (Ahonen, 2009, 2010, 2012) or 10^3 tracks (Bello, 2011; Silva et al., 2013). To date, no large-scale evaluation has been performed. We motivate such an investigation to establish the accuracy of NCD and related discrete-valued approaches for version identification.

We note that our proposed approach of quantifying predictive uncertainty resembles the work of Serrà et al. (2012), who determine similarity by computing the cross-prediction error between feature sequences. Our proposed approach differs in that we aim to evaluate alternative measures to the mean squared error statistic considered in the described work. Serrà et al. consider the role of predictive models. Further, in our own investigation into methods for determining similarity from obtained predictions, we contrast discrete-valued and continuous-valued approaches.

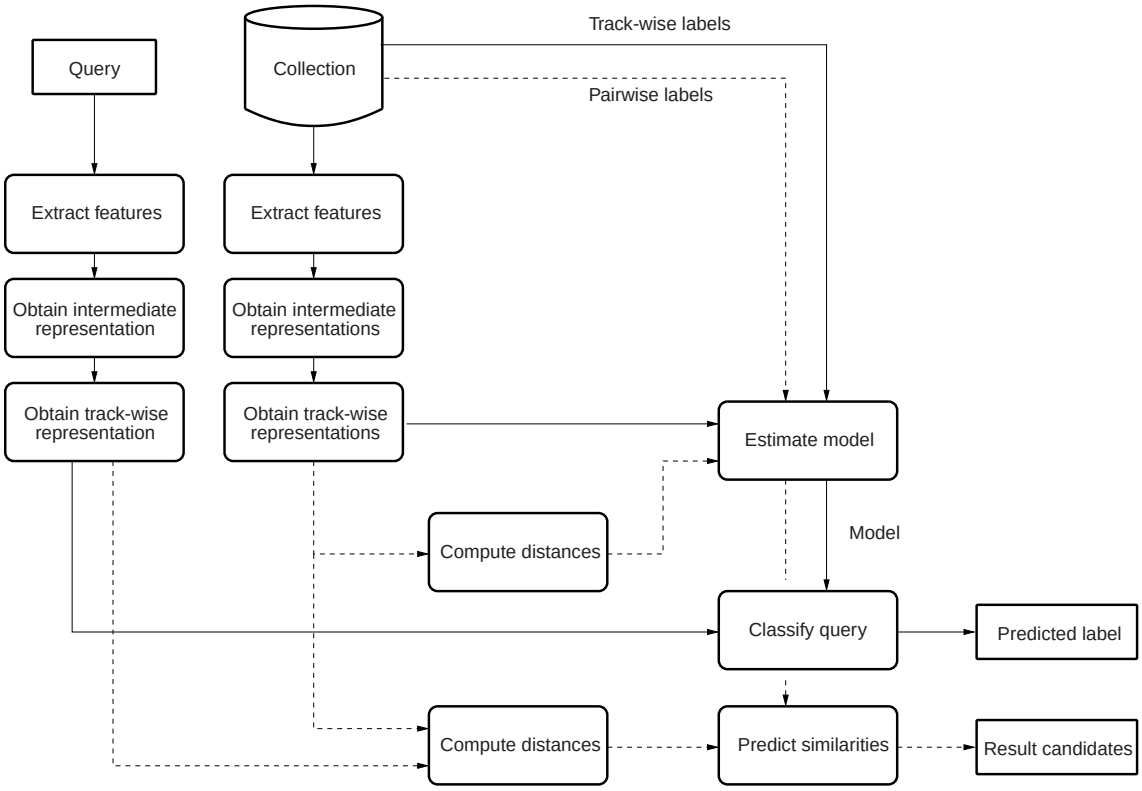


Figure 2.4: Stages involved in a typical system for determining low-specificity similarity. The system incorporates a classifier which is subsequently used to classify tracks. Dashed lines indicate stages for an alternative query-by-example system.

2.4.3 Low-Specificity Similarity

In contrast to version identification, among low-specificity tasks (such as genre classification, artist identification and mood classification) Fu et al. (2011b) observe that a variety of low-level and mid-level features have been considered. Figure 2.4 illustrates the stages involved in a typical system for determining low-specificity similarity. We view Figure 2.4 as a variant of Figure 1.1 with an additional step of obtaining an intermediate feature representation following feature extraction. Similar to version identification, based on sequences of low-level or mid-level features, a number of techniques exist for summarising features, which results in the intermediate representation. Thereafter, obtained sequences may be converted to a track-wise representation in a further step. Finally, the obtained representations may be used for classification; alternatively distances computed between track-wise representations may be used for model estimation and prediction. We base our discussion of the previously described stages on Fu et al. (2011b).

Feature Extraction and Intermediate Representation

Using low-level features, the motivation behind track-wise feature summarisation is to obtain a representation of timbral content which facilitates subsequent application of a similarity model. As described in Section 2.4.1, as low-level features Tzanetakis and Cook (2002) propose the use of MFCCs for genre classification. Such features have been widely adopted; the majority of investigations which we discuss in the following are based on MFCCs. Related features based on weighted magnitude spectra are applied by West et al. (2006); Slaney et al. (2008); Dieleman et al. (2011); Coviello et al. (2012); Dieleman and Schrauwen (2013). Alternative features include scalar-valued statistics obtained from frame-wise magnitude spectra (Tzanetakis and Cook, 2002; Mörchen et al., 2006; Lee et al., 2009). As an alternative to spectral analysis using the STFT, wavelet analysis has also been considered (Li and Ogihara, 2006).

Having obtained a sequence of features for a given track, a straightforward method for obtaining a track-wise feature summary involves computing statistical moments across the entire sequence (Mandel and Ellis, 2005). Chiefly, we may characterise methods based on whether the temporal order of features is discarded or retained (Casey et al., 2008b). The former so-called 'bag-of-features' approach involves estimating distributions of individual observations in a feature sequence (Logan and Salomon, 2001; Aucouturier and Pachet, 2002; Berenzweig et al., 2004; Mandel and Ellis, 2005; Aucouturier et al., 2007; Helén and Virtanen, 2010; Fu et al., 2011a).

The relative convenience of bag-of-features approaches stands in contrast to the importance of temporal structure in perception of musical timbre observed by McAdams et al. (1995). In their study, based on synthesised monophonic instrument sounds the authors perform an exploratory analysis of acoustic correlates of human judgements of timbral similarity. Notably, the authors observe that similarity judgements are explained by log-attack time and spectral flux. Log-attack time quantifies the duration between the onset of a tone and its maximum amplitude. As previously described in Section 2.4.1, spectral flux quantifies changes in magnitude spectrum. On the basis of McAdams et al. (1995), we note that while a bag-of-features approach using such features potentially describes the overall timbral quality in a piece of music, it may fail to capture temporal structure relevant for determining similarity.

Aucouturier et al. (2007) argue that the bag-of-features approach is insufficient to model polyphonic music for determining similarity. Aucouturier et al. evaluate classification accuracy

based on a bag-of-features approach using MFCCs. The collection consists of 350 popular music tracks associated with 37 distinct artists, with tracks selected to maintain timbral homogeneity both within individual tracks and within individual artists. The authors examine the effect of artificially increasing temporal homogeneity in feature sequences, an operation which is performed by constructing audio tracks from random extracts, using a specified extract length and before performing feature extraction. Compared to a baseline classifier using untransformed audio, Aucouturier et al. observe reductions in classification accuracy, as temporal homogeneity increases. As a possible explanation for the reduction in accuracy, the authors posit that the distribution of frames in polyphonic music varies, depending on the considered time scale: Compared to recordings of urban sound-scapes, it is suggested that the distribution of frames in polyphonic music exhibits relatively low amounts of statistical self-similarity. The approach of summarising across the entire set of frames therefore appears insufficient.

One possible approach to mitigating the shortcomings of the bag-of-features approach involves the intermediate step of aggregating features locally, before summarising anew using obtained summary statistics. Tzanetakis and Cook (2002) propose the use of a *texture window*, which captures the local mean and variance of features contained in a 1s window. For the task of predicting musical similarity, Seyerlehner et al. (2010) consider a related approach which uses additional summary statistics. Among statistics, at each window position the authors compute first-order differences and pairwise correlations between feature vectors; the authors consider window sizes in the range of approximately $[0.2\text{s}, 6.0\text{s}]$. Hamel et al. (2011) consider higher-order moments in addition to maximum and minimum feature values, with evaluated window sizes in the range of approximately $[1.5\text{s}, 20\text{s}]$; a window size of 2.3s is observed to be optimal. A similar window size maximises genre classification performance, as observed by Wülfing and Riedmiller (2012). Note that the approach of computing feature differences may also be performed in track-wise manner before windowing, thus a sequence of low-level derivative features is obtained (Mörchen et al., 2006).

An alternative approach characterises the temporal order of features at each window position, by applying spectral analysis. Pampalk (2006, p. 37) proposes fluctuation patterns, which are based on applying a 3s window to frames representing weighted magnitude spectra. For each frequency band and for each window position, Pampalk determines low-frequency modulation of the aforementioned magnitude spectra, with modulation frequencies in the range $[0\text{Hz}, 10\text{Hz}]$.

Based on perceptual considerations, modulation frequencies are subsequently weighted. Thus, a higher-level feature is obtained, where at each window position the dimensions are frequency bands and modulation frequencies. It should be noted that an analogous procedure may be applied to alternative features, such as MFCCs (Pampalk, 2006). Lee et al. (2009) propose further techniques based on modulation spectral analysis of MFCCs and using a window size of approximately 6s.

A further alternative involves estimating a time series model at each window position. With the goal of modelling local temporal structure in MFCCs, Meng et al. (2007) apply a multivariate autoregressive (MAR) model for genre classification. In this model, each feature vector is assumed to be a linear combination of immediately preceding feature vectors, plus a multivariate Gaussian error. The number of preceding feature vectors determines the model's order. Chiefly, an MAR model accounts both for temporal correlation among feature vectors, in addition to correlation among feature vector components. Meng et al. estimate MAR models using window sizes in the range of $[1.2\text{s}, 2.2\text{s}]$. Coviello et al. (2012) propose an alternative approach based on local MAR models for semantic tag classification of songs.

The preceding discussion concerns the use of low-level features for characterising musical timbre. Among mid-level features, those related to musical rhythm have been applied in genre classification and mood identification. Note that some of the previously described methods characterised as low-level also potentially capture rhythmic content, without the subsequent step of tempo estimation (Pampalk, 2006; Seyerlehner et al., 2010). The method proposed by Tzanetakis and Cook (2002) first decomposes an audio signal into frequency bands each with octave bandwidth. For each band, time-domain envelopes are then computed and summed together, before using the sample autocorrelation to detect periodicities in the resulting signal histogram of autocorrelation peaks. This approach has been subsequently applied for the purpose of mood identification (Li and Ogihara, 2006; Yang et al., 2008), in addition to further genre classification approaches (Li and Sleep, 2005; Li and Ogihara, 2006). For the problem of dance rhythm classification, Gouyon et al. (2004) evaluate an extensive set of rhythmic features based on both tempo estimation and periodicity histogram statistics.

Compared to rhythmic features, harmonic features have been applied less widely. Similar to the method for detecting rhythmic periodicities, Tzanetakis and Cook (2002) decompose the audio signal into two frequency bands, before summing time-domain envelopes and computing

the sample autocorrelation. Peak-picking is then applied to the sample autocorrelation, before mapping peaks to pitches, assuming an equal-tempered scale. Tzanetakis and Cook (2002) further fold pitches in a manner similar to a chromagram. A track-wise summary of tonal content is then formed on the basis of folded pitch histograms. In an investigation using beat-synchronous chroma features, Ellis (2007) combines track-wise chroma averages with MFCC feature averages for artist classification.

We note that both low-level and mid-level feature sequences may be quantised before being processed further. A straightforward approach involves vector quantisation using a specified codebook: Foote (1997) applies tree-based quantisation to sequences of MFCCs, whereas Li and Sleep (2005); Reed and Lee (2009); Ren and Jang (2012) use the K-means clustering algorithm for quantisation. Based on the K-means clustering approach, Fu et al. (2011a) extensively evaluate the effect of codebook generation strategies for genre and artist classification. Alternative approaches include hierarchical clustering (Langlois and Marques, 2009) and self-organising maps (Levy and Sandler, 2006). Finally, we may consider chord identification a further approach to quantisation. Using such an approach, Anglade et al. (2010) use chroma features to infer chord sequences.

Track-Wise Feature Representation

As previously mentioned, to obtain a track-wise representation of a feature sequence we may adopt a bag-of-features approach and assume that observations in the entire sequence are independent and identically distributed. For such an approach, we distinguish methods which rely on continuous-valued sequences from those which rely on strings.

Among continuous-valued approaches, a straightforward approach of summarising a sequence involves computing the component-wise mean and variance, as proposed by Tzanetakis and Cook (2002); Levy and Sandler (2006); Lee et al. (2009). To quantify correlation between components, the covariance may be computed (Mandel and Ellis, 2005). We may interpret both approaches as estimating the parameters of a multivariate Gaussian distribution. Alternatively, a Gaussian mixture model (GMM) may be estimated; the sequence is thus parametrised by a number of multivariate Gaussians, in addition to a weighting vector. In a GMM, the probability density of an observation is the weighted combination of probability densities associated with each estimated Gaussian. Aucouturier and Pachet (2002); Tzanetakis and Cook (2002) propose a 3-component model, whereas Mandel and Ellis (2005) estimate 20 components. In contrast

to the single Gaussian model which may be estimated using descriptive statistics, estimating a GMM requires iterative algorithms such as K-means clustering (cf. Berenzweig et al., 2004) and Expectation-Maximisation (cf. Helén and Virtanen, 2010). We note that besides computing the component-wise mean and variance, a number of alternative summary statistics may be computed: Mörchen et al. (2006); Seyerlehner et al. (2010) compute higher-order moments and percentiles, respectively. Note that all described continuous-valued approaches may be used to summarise a track using a vector whose dimensionality is invariant to song duration.

Among discrete-valued approaches, as previously described for the case of mid-level rhythmic features (cf. Tzanetakis and Cook, 2002) one possible approach to summarising features involves computing a histogram. For the case of MFCCs (Foote, 1997; Logan and Salomon, 2001; Levy and Sandler, 2006; Fu et al., 2011a), this typically involves a non-trivial quantisation algorithm such as K-means clustering, contrasting with uniform partitioning used for the case of autocorrelation peaks (Tzanetakis and Cook, 2002). In either case, similar to the aforementioned continuous-valued approaches, the dimensionality of histograms is invariant to the duration of songs. If histogram counts are subsequently normalised to sum to one, the normalised counts are sample estimates of event probabilities associated with histogram bins. In this way, the histogram is a non-parametric model of the marginal distribution of observations in the feature sequence.

In both described discrete-valued and continuous-valued approaches, the obtained representations are used for subsequent similarity modelling. We note that such representations are not used exclusively: in a further bag-of-features approach, models are directly applied to pooled sequences considered as sets of observations (cf. Berenzweig et al., 2004). Other models account for temporal structure in sequences (cf. Li and Sleep, 2005).

Similarity Models

One class of similarity model involves pairwise comparisons between tracks. As a bag-of-features approach, Aucouturier and Pachet (2002) model MFCC features using GMMs, so that each song in a collection is represented as a GMM. To compare the pair of songs (i, j), a set of samples is obtained with respect to the model GMM_i. The cross-likelihood of the obtained samples using the model GMM_j is then computed. In this manner, cross-likelihood is used as a pairwise similarity measure between tracks for music recommendation and playlist generation tasks. A widely-applied distance measure is the Kullback-Leibler divergence (KLD), which may be estimated using sampling techniques for GMMs; it may be computed in closed form for mul-

tivariate Gaussians and discrete distributions (cf. Helén and Virtanen, 2010). The closed-form approach using single Gaussians is considered by Levy and Sandler (2006); the closed-form approach using histograms is considered by Vignoli and Pauws (2005); Fu et al. (2011a). Using GMMs, Aucouturier et al. (2007) estimate the KLD using sampling techniques. Logan and Salomon (2001); Berenzweig et al. (2004) evaluate alternative distance functions which may be computed exactly, given obtained GMM parameters. Besides computing pairwise similarities between tracks, Berenzweig et al. (2004) estimate GMMs using features pooled across individual artists. In this way, the authors quantify pairwise similarity between artists.

Pairwise distances may be used to classify tracks. A comparatively straightforward approach involves K-nearest neighbours (KNN) classification. Here, training data are represented as a set of exemplars with associated labels. Given a query track, a set of nearest neighbours is determined in the set of exemplars, using a specified distance measure. The query is then classified by applying a specified voting function to the set of nearest neighbours, such as a majority vote.

Tzanetakis and Cook (2002); Li and Ogihara (2006); Lee et al. (2009) apply KNN for genre classification, with the number of nearest neighbours in the range $[1..10]$. Among distance functions, Mandel and Ellis (2005) evaluate the KLD for artist classification, representing tracks as single Gaussians. Where tracks may be summarised as vectors of constant dimensionality (as is the case for single Gaussians), alternative distance measures include the Euclidean (cf. Fu et al., 2011a) and Mahalanobis distances (cf. Mandel and Ellis, 2005). Fu et al. (2011a) evaluate distances between histograms, including the KLD. In the work of Slaney et al. (2008), supervised methods are employed which learn a linear transformation of features for KNN classification based on the Euclidean distance. As stated previously, use of the KNN algorithm does not preclude using non-vector representations: for example, the KLD may be estimated using sampling techniques (Mandel and Ellis, 2005).

Expanding on the use of pairwise similarity between tracks for classification, support vector machines (SVMs, cf. Cristianini and Shawe-Taylor, 2000) have been applied extensively in genre and artist classification (Li and Sleep, 2005; Mandel and Ellis, 2005; Meng and Shawe-Taylor, 2005; Li and Ogihara, 2006; Mörchen et al., 2006; Meng et al., 2007; Reed and Lee, 2009; Coviello et al., 2012; Ren and Jang, 2012; Wülfing and Riedmiller, 2012). In SVMs, classification is chiefly performed with respect to a hyperplane in high-dimensional space. The hyperplane is determined by optimising a measure of separation between classes in the training

data. Importantly, the feature space used for optimisation need not conform to the original feature space. By mapping the original feature space to higher-dimensional space, finding the optimal hyperplane is facilitated when the training data are not linearly separable. Notably, while SVMs are intrinsically linear classifiers, use of appropriate *kernel* functions allows SVMs to behave as non-linear classifiers with respect to the original feature space. A kernel function is conceptually a pairwise similarity function; it computes the inner product between two objects represented in the expanded feature space.

In a bag-of-features approach, Mandel and Ellis (2005) transform the KLD for use as an SVM kernel. The authors perform artist classification by training an SVM using artist annotations as class labels. They compute track-wise feature averages, thus each training datum corresponds to a single track alongside an artist label. Likewise, a linear kernel is defined as the dot product between feature vectors, thus it is possible to use this approach to compare track-wise feature averages (Meng and Shawe-Taylor, 2005; Li and Ogihara, 2006). Meng and Shawe-Taylor (2005) propose the use of alternative kernels: the evaluated convolution kernel relies on exhaustively computing pairwise correlations between feature vectors, and thus does not require feature summarisation. The evaluated product probability kernel is a divergence measure similar to the KLD, yet it admits a closed-form solution when comparing GMMs. Coviello et al. (2012) propose further kernel functions which are suited for comparing mixtures of MAR models.

Contrasting with SVMs, in other classifiers a pairwise similarity function is not specified explicitly. Tzanetakis and Cook (2002); Li and Ogihara (2006); Ellis (2007) estimate the distribution of summary feature vectors in each class using a GMM. Query tracks are then classified by determining the GMM most likely to have emitted summary vectors associated with queries. A further approach involves the use of classification trees (West et al., 2006; Foucard et al., 2011). As proposed by West et al., rather than produce a single prediction for a query track, predicted genre labels are obtained for each feature vector associated with the query track. Thus a histogram of predicted genres is obtained. The authors propose that normalised genre histograms are compared. In this way, the output of a classifier may be used to quantify pairwise similarity between tracks.

Modelling Temporal Structure

The similarity models previously discussed are bag-of-features approaches. Among methods which attempt to model temporal structure, we distinguish between methods which rely on

continuous-valued sequences and methods which rely on strings.

As discussed in Section 2.3.3, one possible discrete-valued approach involves Markov models estimated from n-grams. Langlois and Marques (2009) quantise sequences of MFCCs using a two-stage procedure, where in the first stage track-wise summaries are obtained by estimating a GMM for each track. The obtained GMMs are then used to estimate a global codebook using K-means clustering. The codebook is used to obtain bi-grams and estimate a first order Markov model. For the purpose of genre classification, Langlois and Marques estimate Markov models across individual genres and compute cross-likelihoods between query sequences and each genre model. Langlois and Marques further estimate artist-wise and track-wise models for artist classification and similarity estimation for playlist generation, respectively. For semantic tag classification, Reed and Lee (2009) propose to segment sequences of MFCCs, before quantising segment centroids using an iterative procedure and computing uni-gram and bi-gram distributions. The obtained distributions are then classified using a set of SVMs, with each SVM trained to predict a specific tag.

For genre classification, Ren and Jang (2012) propose a similar combination of feature segmentation and quantisation to Reed and Lee (2009), which results in a discrete-valued transcript for each track. However, instead of using transcripts to estimate Markov models, Ren and Jang apply a heuristic to transcripts, with the goal of identifying a set of characteristic sub-sequences for each genre. Sub-sequences may originate from non-contiguous events within transcripts, thus facilitating efficient representation of long-term repetition structure. Similar to Reed and Lee (2009), distributions of characteristic sub-sequences are thereafter classified using an SVM.

Li and Sleep (2005) propose a similarity measure between strings for genre classification. The similarity measure is based on the NCD, which as described in Section 2.4.2 quantifies pairwise similarity using a string compressor. Li and Sleep propose a modification to the LZ compressor (Ziv and Lempel, 1978), which generates a histogram of substring occurrences in a given string. Similarity between strings is then determined by computing the inner product between histograms obtained for each string. As their features, Li and Sleep use MFCCs, which they subsequently quantise using the K-means algorithm. To perform classification, the authors apply their similarity measure as an SVM kernel.

A further discrete-valued approach proposed by Anglade et al. (2010) uses chord sequences as a high-level representation of harmonic content. For the purpose of genre classification,

Anglade et al. apply inductive logic programming (Muggleton, 1991) to a corpus of annotated chord sequences. The authors in this way attempt to identify genre-characteristic chord sequences. Analogous to Ren and Jang (2012), the learnt representations may originate from non-contiguous chord sequences in the training data. Thus, irrelevant chords in a characteristic chord sequence may be ignored. Inductive logic programming yields a set of rules by which a chord sequence may be classified. Anglade et al. apply chord transcription to classify tracks from audio, using the set of learnt rules.

Recent approaches attempt to model temporal structure using representations constructed at multiple time scales. Based on a bag-of-features approach and using low-level features, Foucard et al. (2011) propose an ensemble of classifiers, where each classifier is trained on a windowed sequence of features at a given time scale, with label annotations supplied for each track. Representations at successive resolutions are obtained by aggregating feature frames using averaging. Each classifier is a decision tree, whose nodes specify feature thresholds used to form predictions. Applied to the task of semantic tag classification, a given query track is classified by combining predictions obtained at each window position, at each time scale.

Dieleman and Schrauwen (2013) propose to learn multi-scale feature representations. To this end, they compute short-time spectra, using feature averaging schemes similar to Foucard et al. (2011) to obtain spectra at multiple time scales. Representations are subsequently obtained by decorrelating feature vectors, then multiplying resulting vectors with a dictionary matrix. Use of spherical K-means clustering to learn the dictionary matrix is proposed following Coates and Ng (2012). This approach has shown potential as an unsupervised method for sparse coding, whereby given input vectors are closely approximated as linear combinations of basis vectors in a dictionary matrix, subject to the constraint that the associated weighting vectors are sparse. The obtained representations are aggregated across a window; each window is then used to obtain predicted semantic tags, based on an ANN.

Finally, deep neural network architectures have been proposed for modelling temporal structure. A deep network may be defined as a feed-forward network consisting of many hidden layers, with each layer an RBM (cf. Section 2.3.3; Hinton, 2007). Dieleman et al. (2011) apply a two-stage learning process. In the first stage, a deep architecture is trained in unsupervised fashion (Hinton, 2002). In the employed convolutional architecture, connections between layers are constrained to enforce locality (each unit at level i is connected to a small number of proximate

units at level $i-1$) and translation invariance (at levels i, $i-1$, if there is a connection between units j, k and between units $j+c$, $k+c$ at respective levels, the connection weights are identical). Additional *pooling layers* aggregate unit activations by computing maxima. Together with the use of pooling layers, the constraints of locality and translational invariance allow successive layers to learn successively higher-level representations. Having trained the network in unsupervised fashion, Dieleman et al. employ a supervised learning stage based on backpropagation (Bryson et al., 1963). The resulting network is applied to artist and genre classification, using beat-aligned chroma and timbral features. Hamel et al. (2011) propose a similar hierarchical architecture for semantic tag classification, investigating the role of pooling functions.

Large-Scale Approaches

Investigations on low-specificity similarity for retrieval using large collections have focussed on summary feature representations. As described in Section 2.4.2 for cover song identification, such representations permit the use of computationally efficient methods for nearest neighbour retrieval with respect to a query. In contrast to version identification, the retrieval process dispenses with subsequent pairwise comparisons between feature sequences.

Schnitzer et al. (2009) note that despite its widespread use as a pairwise distance measure, the KLD potentially presents a challenge when applied to large datasets. Firstly, it does not fulfil metric requirements: while it is non-negative and symmetric and while it may be trivially symmetrised, it does not fulfil the triangle inequality. As a result, nearest-neighbour retrieval requires a linear scan of items in a collection, an approach which may not be sufficiently scalable. Secondly, computing the KLD requires a relatively large number of floating point operations, compared to alternative distances such as the Euclidean metric. To address the latter issue, Schnitzer et al. propose the use of FastMap (Faloutsos and Lin, 1995), a method for mapping objects to Euclidean space, given a pairwise distance matrix between objects. Notably, FastMap empirically has been shown to perform well, even if a restricted set of pairwise distances is available, rather than a pairwise distance matrix for the entire collection considered. Thus, Schnitzer et al. use FastMap to map track-wise single Gaussian representations of MFCCs to Euclidean space. Given a query, the authors obtain nearest neighbours in Euclidean space using a linear scan. The authors propose a filter-and-refine strategy, whereby nearest neighbours are subsequently re-ranked using the KLD. Based on a collection of 2.5 million tracks, Schnitzer et al. report retrieval times of approximately 0.5s per query, while retaining 95% recall of nearest neighbours with respect

to the original feature space.

Schlüter (2013) applies alternative methods for efficient nearest-neighbour retrieval. Using the same dataset as Schnitzer et al. (2009), Schlüter contrasts methods which provide sub-linear nearest-neighbour retrieval complexity, against methods which aim at performing pairwise comparisons at low computational cost while performing a linear scan of the collection. The evaluation is based on applying a filter-and-refine strategy to the similarity measures proposed by Seyerlehner et al. (2010), Mandel and Ellis (2005). As a measure of performance, Schlüter computes execution time subject to the requirement of 90% nearest neighbour recall using a full linear scan with the original similarity measures. Among filtering methods with sub-linear retrieval, in addition to LSH Schlüter applies linear and non-linear mappings of features to binary vector representations; akin to LSH such representations allow for $O(1)$ retrieval using lookup tables. Among filtering methods based on linear scans, besides FastMap Schlüter applies PCA and ANNs for dimensionality reduction, the resulting vectors are subsequently compared using the Euclidean distance. As observed, while sub-linear approaches provide better scalability, for the evaluated collection on the scale of 10^6 tracks, linear-time filtering techniques when applied to high-dimensional features yield superior speedup.

Discussion

In the preceding, we have reviewed existing approaches to determining low-specificity similarity. Concerning methods for obtaining track-wise feature representations, we may distinguish between bag-of-features representations and track-wise representations which account for temporal structure in sequences.

The extent to which the reviewed methods account for temporal structure in music demands more refined distinction among approaches. Above the level of extracted feature frames, a variety of approaches model local temporal structure while discarding temporal structure in the resulting sequence when computing the track-wise summary (Tzanetakis and Cook, 2002; Meng and Shawe-Taylor, 2005; Li and Ogihara, 2006; Pampalk, 2006; Meng et al., 2007; Lee et al., 2009; Seyerlehner et al., 2010). A further possibility involves accounting for both local and global temporal structure, as exemplified by multi-scale approaches (Dieleman et al., 2011; Foucard et al., 2011; Dieleman and Schrauwen, 2013).

A particular category of track-wise representation yields equal-sized vectors, such as may be obtained by computing histograms, summary statistics or by estimating single Gaussians. With a

view to performing large-scale retrieval, vector-valued representations facilitate computationally efficient pairwise comparisons. However, the described methods for obtaining such representations typically discard temporal structure. As discussed in Section 2.3.4, we propose to compute information-theoretic measures of predictability as summary statistics.

Among summary statistics which account for temporal structure in feature sequences, we note the work of Mörchen et al. (2006), who performs a large-scale evaluation on the utility of summary statistics including autocorrelation and partial autocorrelation coefficients (cf. Lütkepohl, 2005), out of a total of 164 summary statistics. However, the authors do not consider any information-theoretic measures of predictability. Similarly, we note that while Streich (2006) considers the possibility of using information-theoretic measures of predictability, no evaluations are performed using the approach.

2.5 Conclusion

In this chapter, we have reviewed computational methods for determining musical similarity. The background for our approach is the concept of musical expectation. Musical expectation plays an important role in the cognition of musical structure, the latter which in turn is relevant in determining musical similarity. While we do not aim at simulating cognitive processes for our purposes in music content analysis, we motivate our own approach having considered existing models of musical expectation.

In particular, we have identified information-theoretic measures of predictability as a means of quantifying regularity in musical sequences. Among existing approaches, there are methods applicable to both continuous-valued and discrete-valued sequences. Our working hypothesis is that measures of predictability might be used to determine music with similar temporal structure; we deem sequences musically similar, if they incur similar amounts of predictability. An information-theoretic approach offers the properties of abstraction, generality and expressiveness.

Considering potential applications for our information-theoretic approach, we have identified version identification and low-specificity similarity tasks. For version identification, we seek to compare our approach to existing discrete-valued and continuous-valued methods for quantifying pairwise predictability between tracks. For low-specificity similarity tasks, we seek to evaluate the utility of our approach as a track-wise summary statistic.

In the following chapter 3 we review information-theoretic measures of predictability and detail the methods used in our investigations.

Chapter 3

Information-Theoretic Methods

In this chapter, we review information-theoretic concepts relevant to our investigations in chapters 4 and 5.

We begin in Section 3.1 by reviewing Shannon's information theory; sections 3.1.1 and 3.1.2 respectively discuss measures of predictability on discrete-valued and continuous-valued memoryless sources. Section 3.1.3 reviews analogous measures applicable to information sources with memory.

In Section 3.2, we review the concept of algorithmic information content, which we may view as a deterministic measure of information contained in a mathematical object, such as a string. In Section 3.2.2 we review relations between algorithmic information content and Shannon information. In Section 3.2.3 we review the normalised compression distance, a measure of similarity between strings. Section 3.3 briefly discusses the interpretation of our considered information-theoretic measures as measures of predictability. Finally, in Section 3.4, we conclude our discussion.

3.1 Shannon Information

For an extensive discussion of information-theoretic concepts reviewed in the following, we refer to MacKay (2003); Cover and Thomas (2012). Our own discussion closely follows Feldman (2002), who provides a more concise overview.

3.1.1 Discrete Random Variables

We denote with X, Y random variables whose sample spaces are both given by a finite alphabet $\mathcal{A}$. We denote with $P_X(x)$ the probability of observing x, with $x \in \mathcal{A}$. We denote with $P_{X,Y}(x,y)$ the joint probability of observing x and y, with $y \in \mathcal{A}$. Finally, we denote with $P_{X|Y}(x|y)$ the conditional probability of observing x, given the observation y. In a sequence of random variables $X_{1:N} = (X_1, X_2, \ldots, X_N)$, we use $P_{X_{1:N}}(x_{1:n})$ to denote the joint probability of observations $x_{1:n} = (x_1, x_2, \ldots, x_N)$, with $x_{1:N} \in \mathcal{A}^N$.

The fundamental information-theoretic quantity for a single random variable X is the entropy $H(X)$, defined as

$$H(X) = -\sum_{x \in \mathcal{A}} P_X(x) \log P_X(x). \tag{3.1}$$

A straightforward interpretation of Equation 3.1 is that $H(X)$ quantifies uncertainty about outcomes in terms of their respective probabilities: for a given alphabet size, entropy is maximised for uniformly distributed outcomes, whereas entropy is minimised for distributions whose probability mass is assigned to a single outcome. Defining with $\mathcal{L}_X(x)$ the surprisingness of observation x,

$$\mathcal{L}_X(x) = -\log P_X(x) \tag{3.2}$$

we note that $H(X)$ is the expectation value $\mathbf{E}[\mathcal{L}_X]$. Thus, we may alternatively interpret entropy as the average surprisingness of the observed outcome of X. Taking the logarithm to base 2, entropy is measured in bits.

The entropy $H(X)$ is non-negative and continuous with respect to the distribution of outcomes in X. Since entropy is a function of the distribution of outcomes alone, it does not depend on the actual values of outcomes. It follows that entropy is invariant to any re-arranging of outcomes assigned to probability values.

The joint entropy $H(X,Y)$ is defined as

$$H(X,Y) = -\sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{A}} P_{X,Y}(x,y) \log P_{X,Y}(x,y). \tag{3.3}$$

Equation 3.3 quantifies in analogous manner to $H(X)$ the uncertainty about pairs of outcomes in the random variables X, Y. For independent X, Y, we have $H(X,Y) = H(X) + H(Y)$.

We may re-state $H(X,Y)$ as the entropy $H(Z)$ of the random variable $Z = (X,Y)$ whose sample space is $\mathcal{A}^2$ and whose probability mass function is $P_Z = P_{(X,Y)}$. It follows that we may extend the definition of joint entropy to a sequence of random variables $X_{1:N} = (X_1, X_2, \ldots, X_N)$,

for which the joint entropy $H(X_{1:N})$ is defined as

$$H(X_{1:N}) = -\sum_{x_{1:N}\in\mathcal{A}} P_{X_{1:N}}(x_{1:N}) \log P_{X_{1:N}}(x_{1:N}). \tag{3.4}$$

Finally, the conditional entropy $H(X|Y)$ is defined as

$$H(X|Y) = -\sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{A}} P_{X,Y}(x,y) \log P_{X|Y}(x|y) \tag{3.5}$$

$$= H(X,Y) - H(Y) \tag{3.6}$$

The conditional entropy $H(X|Y)$ quantifies the uncertainty about outcomes in X, given knowledge of outcomes in Y. For sequences of random variables $X_{1:N}, Y_{1:M}$, it follows that we may express the conditional entropy $H(X_{1:N}|Y_{1:M})$ as

$$H(X_{1:N}|Y_{1:M}) = H(X_{1:M}, Y_{1:M}) - H(Y_{1:M}). \tag{3.7}$$

Shannon (1948) proposes entropy in the context of an information source which transmits a sequence of messages via a communication channel to a receiver. A *discrete source* is a sequence of discrete random variables $\underline{X} = (X_1, X_2, \ldots)$ with respective outcomes in $\mathcal{A}$. We define with $H_\mu(\underline{X})$ the entropy rate,

$$H_\mu(\underline{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_{1:n}) \tag{3.8}$$

which we may interpret as the average uncertainty about outcomes, while accounting for correlation among observations. In a *discrete memoryless source*, random variables in the sequence are independent and identically distributed. Based on Equation 3.3 and owing to the independence and identical distribution of random variables, for a discrete memoryless source we have $H(X_{1:N}) = NH(X_1)$; it follows that $H_\mu(\underline{X}) = H(X_1)$.

For any source, we may devise a code which maps a sequence of random variable outcomes to a sequence of binary codewords; we may use such a code to compress data for transmission. We require that the code is uniquely decodable, i.e. the code must yield a lossless representation of the original sequence of observations. For the case of a discrete memoryless source, an efficient approach involves constructing a code which maps each outcome to a codeword, while accounting for the respective marginal probabilities of outcomes. Preferably, probable outcomes should map to short binary strings, whereas improbable outcomes should map to long binary strings. What is the bound on achievable compression? Intuitively stated, Shannon's source coding theorem states that for a single observation, the expected codeword length L in bits is

bounded as

$$L \geq H_\mu(\underline{X}) \tag{3.9}$$

with logarithms in $H_\mu(\underline{X})$ taken to base 2. If we assume that we have an optimal code, it follows that we may interpret the entropy $H(X)$ of a random variable as the expected number of bits required to represent its value.

Thus, the notion of 'information' in entropy as proposed by Shannon is in terms of the distribution of possible content in a given message.

Given random variables X, Y, we may quantify disparities in uncertainty between the respective variables. The cross entropy $H^\times(X,Y)$ is defined as

$$H^\times(X,Y) = -\sum_{x \in \mathcal{A}} P_X(x) \log P_Y(x). \tag{3.10}$$

If we recall the definition of $H(X)$ in Equation 3.1, we may interpret $H^\times(X,Y)$ as the expected number of bits required to represent the value of X, given an optimal code for Y. A further measure of disparity between X, Y is the Kullback-Leibler divergence (KLD) $D_{\mathrm{KL}}(P_X \| P_Y)$, defined as

$$D_{\mathrm{KL}}(P_X \| P_Y) = \sum_{x \in \mathcal{A}} P_X(x) \log \frac{P_X(x)}{P_Y(x)} \tag{3.11}$$

$$= H^\times(X,Y) - H(X). \tag{3.12}$$

From our preceding discussion, we may interpret the KLD as the expected number of additional bits required to represent the value of X, given an optimal code for Y.

3.1.2 Continuous Random Variables

The information-theoretic measures discussed are for discrete random variables. Next, we review analogous measures for continuous random variables. Henceforth, we denote with X, Y continuous random variables whose respective real-valued outcomes can be described in terms of probability density functions. We denote with $p_X(x)$ the probability density associated with observing x. We denote with $p_{X,Y}(x,y)$ the joint probability density of observing x and y. Further, we denote with $p_{X|Y}(x|y)$ the conditional probability density of observing x, given the observation y.

The continuous entropy $h(X)$ is defined as

$$h(X) = -\int p_X(x) \log p_X(x)\, dx. \tag{3.13}$$

The continuous joint entropy $h(X_{1:N})$ and conditional entropy $h(X_{1:N}|Y_{1:M})$ are respectively defined as

$$h(X_{1:N}) = -\int p_{X_{1:N}}(x_{1:N}) \log p_{X_{1:N}}(x_{1:N})\, dx_{1:N} \tag{3.14}$$

and

$$h(X_{1:N}|Y_{1:M}) = h(X_{1:N}, Y_{1:M}) - h(Y_{1:M}). \tag{3.15}$$

Compared to Equations 3.1, 3.4, 3.7, we might conjecture plausibly that the continuous entropies $h(X)$, $h(X_{1:N})$, $h(X_{1:N|Y_{1:M}})$ share properties with their discrete-valued counterparts. Indeed, continuous entropy is continuous with respect to the probability density functions under consideration. For continuous X, we may interpret $h(X)$ in similar manner as a measure of uncertainty to $H(X)$: continuous entropy is maximised for a uniform distribution, given a specified range for outcomes in X (Shannon, 1948).

However, as suggested from the latter requirement of specifying a range for outcomes in X, continuous entropy quantifies uncertainty with respect to a given coordinate system (Shannon, 1948). Moreover, continuous entropies may be negative, since for a given uniform distribution there exists a coordinate system under which the entropy is zero (cf. Cover and Thomas, 2012).

Clearly, the interpretation of continuous entropy as the expected number of bits required to represent a random variable does not hold, since a real number may be represented with arbitrary accuracy. Despite this caveat, we may nevertheless adopt a comparable interpretation if we specify a required accuracy. This interpretation is based on integration in the Riemannian sense, whereby the domain of the integrand is quantised using a given bin width δ. Integration is then performed as a summation in the limit as $\delta \to 0$. We define with $H(X_\delta)$ the discrete entropy of the continuous random variable X quantised using bin size δ. Providing X is Riemann integrable (i.e. the probability density is bounded and continuous with respect to a given compact interval), it may be shown that for small δ (cf. Cover and Thomas, 2012),

$$H(X_\delta) \approx h(X) - \log_2 \delta. \tag{3.16}$$

Setting $\delta = 2^{-n}$, we obtain

$$H(X_\delta) \approx h(X) + n. \tag{3.17}$$

From Equation 3.17 we may interpret n as the expected number of bits required to represent the value of a uniformly distributed continuous random variable to which we apply n-bit quantisation.

$h(X)$ is then the expected number of bits required to represent the value of the quantised random variable X, minus an offset for n-bit quantisation accuracy.

Given continuous random variables X, Y, we may quantify disparities in uncertainty analogous to the discrete-valued measures previously discussed. The cross entropy $h^{\times}(X,Y)$ is defined as

$$h^{\times}(X,Y) = -\int p_X(x) \log p_Y(x)\, dx. \tag{3.18}$$

Similarly, the KLD $D_{\mathrm{KL}}(p_X \| p_Y)$ is defined as

$$D_{\mathrm{KL}}(p_X \| p_Y) = \int p_X(x) \log \frac{p_X(x)}{p_Y(x)}\, dx. \tag{3.19}$$

Similar to $H^{\times}(X,Y)$, we interpret $h^{\times}(X,Y)$ as the expected number of bits required to represent the value of X, given an optimal code for Y, minus a constant number of bits according to the specified accuracy of distinguishing among outcomes in X, Y. For the case of continuous KLD, note that we have

$$\begin{aligned} D_{\mathrm{KL}}(p_X \| p_Y) &= h^{\times}(X,Y) - h(X) && (3.20)\\ &\approx H^{\times}(X_\delta, Y_\delta) - \log_2 \delta - (H(X_\delta) - \log_2 \delta) && (3.21)\\ &= H^{\times}(X_\delta, Y_\delta) - H(X_\delta). && (3.22) \end{aligned}$$

The continuous and discrete KLD therefore have identical interpretation as the expected number of additional bits required to represent the value of X, given an optimal code for Y. In contrast to the preceding continuous measures, the KLD is invariant to the considered coordinate system and is non-negative (cf. Cover and Thomas, 2012).

Considering their similar interpretation and analogous identities, in the following we use capital letters H to refer interchangably to discrete and continuous information quantities. Lowercase letters h to refer exclusively to continuous information quantities.

3.1.3 Sources with Memory

The source coding theorem extends to sources with finite memory, where there may exist n, m such that $X_n, X_m \in \underline{X}$ are correlated. For such processes the expected codeword length required to represent a single observation X_n is—as before—bounded by $H_\mu(\underline{X})$. In contrast to memoryless sources, for sources with memory we have $H_\mu(\underline{X}) \leq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i)$. Thus, there may exist redundancies among observations.

Recall that $H_\mu(X)$ is defined as

$$H_\mu(\underline{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_{1:n}). \tag{3.23}$$

Given two sources $\underline{X}, \underline{Y}$, akin to the joint entropy $H(X,Y)$ we define the joint entropy rate $H_\mu(\underline{X},\underline{Y})$ as

$$H_\mu(\underline{X},\underline{Y}) = \lim_{n\to\infty} \frac{1}{n} H((X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)) \tag{3.24}$$

$$= \lim_{n\to\infty} \frac{1}{n} H(X_{1:n},Y_{1:n}). \tag{3.25}$$

We interpret $H_\mu(X,Y)$ as the average uncertainty in a single pair of observations (X_n,Y_n), accounting for correlation between sources and across successive observations. Further, akin to the conditional entropy $H(X,Y)$, we define the conditional entropy rate $H_\mu(\underline{X},\underline{Y})$ as

$$H_\mu(\underline{X}|\underline{Y}) = H_\mu(\underline{X},\underline{Y}) - H_\mu(\underline{Y}). \tag{3.26}$$

We interpret $H_\mu(\underline{X}|\underline{Y})$ as the average uncertainty in a single observation X_n, accounting for correlation among observations emitted by $\underline{X}$ and given knowledge of observations emitted by $\underline{Y}$.

In Chapter 4, we further consider the cross entropy rate $H^\times_\mu(\underline{X},\underline{Y})$, for discrete random variables defined as

$$H^\times_\mu(\underline{X},\underline{Y}) = \lim_{n\to\infty} -\frac{1}{n} \sum_{x_{1:n}\in\mathcal{A}} P_{X_{1:n}}(x_{1:n}) \log P_{Y_{1:n}}(x_{1:n}). \tag{3.27}$$

We interpret $H^\times_\mu(\underline{X},\underline{Y})$ as the expected number of bits required to represent a single observation emitted by source $\underline{X}$, given an optimal code for source $\underline{Y}$ and accounting for correlation among observations emitted by $\underline{X}$.

Given an empirical estimate $\hat{P}_{X_{1:N}}$ of the distribution $P_{X_{1:N}}$, such as we might obtain by applying a compression algorithm to the sequence $x_{1:N}$, we may estimate $H_\mu(\underline{X})$ using the average log-loss $\ell(\hat{P}_{X_{1:N}}, x_{1:N})$ (cf. Begleiter et al., 2004), defined as

$$\ell(\hat{P}_{X_{1:N}}, x_{1:N}) = -\frac{1}{N} \log \hat{P}_{X_{1:N}}(x_{1:N}). \tag{3.28}$$

In Equation 3.28, taken to base 2 the term $-\log \hat{P}_{X_{1:N}}(x_{1:N})$ corresponds to the number of bits required to represent the sequence of observations $x_{1:N}$, given the estimated model $\hat{P}_{X_{1:N}}$. Dividing by the number of observations N, we may take the observation-wise average number of bits as an estimate of $H_\mu(\underline{X})$. As described in Chapter 4 we estimate $H^\times_\mu(\underline{X},\underline{Y})$ analogously, using an empirical estimate $\hat{P}_{Y_{1:N}}$ of the distribution $P_{Y_{1:N}}$. We then compute $\ell(\hat{P}_{Y_{1:N}}, x_{1:N})$, where $x_{1:N}$ is a sequence of observations emitted by $\underline{X}$.

Markov Sources

Markov sources form a particular class of source with memory, in which the conditional distribution of observation X_n given previous observations $X_{1:n-1}$ is equal to the conditional distribution given context $X_{n-k:n-1}$. That is,

$$P_{X_n|X_{1:n-1}} = P_{X_n|X_{n-k:n-1}} \tag{3.29}$$

where k is the order of the Markov source and where $n > k$. A Markov source is stationary if the conditional distribution of X_n given $X_{n-k:n-1}$ is invariant with respect to n,

$$P_{X_{n+1}|X_{n-k+1:n}} = P_{X_n|X_{n-k:n-1}}. \tag{3.30}$$

We may attain favourable compression performance when applying widespread compression algorithms to stationary, finite-order Markov sources: for such sources, the Lempel-Ziv (LZ) (Ziv and Lempel, 1977) algorithm closely approximates the optimal codeword length $H_\mu(\underline{X})$ when applied to sufficiently long sequences, using a sufficiently large window size (cf. Gallager, 2008). Alternative algorithms such as prediction by partial matching (PPM) (Cleary and Witten, 1984) assume a Markov source.

3.2 Algorithmic Information Content

For an extensive discussion of concepts reviewed in the following, we refer to Li and Vitányi (2008). Our own discussion closely follows Grünwald and Vitányi (2004), who provide a more concise overview.

3.2.1 Information Measures

Informally stated, the algorithmic information content (AIC, alternatively Kolmogorov complexity) of a string is the length of the shortest program which outputs the string under consideration, before terminating. Thus, in AIC the notion of information is in terms of obtaining the ultimate compressed representation of an object; we may conceptualise the object in question as the content in a given message which we seek to encode. This contrasts with Shannon information, where the notion of information is in terms of the distribution of outcomes in a random variable; we may conceptualise such outcomes as possible content in a given message.

More formally, we denote with $\{0,1\}^*$ the set of finite binary strings. We denote with $|x|$ the length of the string $x \in \{0,1\}^*$. The set $\{0,1\}^*$ contains the empty string ε, for which we have

$|\varepsilon| = 0$. Assume that we have a reference machine U, namely a universal Turing machine which when provided with a self-delimiting representation of the pair $z = \langle i, y \rangle$, before halting outputs $U(z)$, the value of the ith computable function with argument y. This assumes that computable functions are enumerable (Turing, 1936). We may thus interpret i as an encoding of a Turing machine T_i whose input is y; the pair $\langle i, y \rangle$ specifies a program whose output is $U(z)$.

The AIC $K(x)$ of the string x may thus be defined as

$$K(x) = \min_z \{|z| : U(z) = x, z \in \{0,1\}^*\} \tag{3.31}$$

which we may interpret as the length in bits of the shortest program which outputs x and then halts. It may be shown that up to an additive constant, $K(x)$ is equivalent when defined in terms of alternative, equally powerful reference machines: alternatively, we may interpret $K(x)$ in terms of the shortest program in a given language such as Java or C which outputs x and then halts. A string x is *algorithmically random* if the length of its shortest program is greater than or equal to the length of x.

In addition to $K(x)$, we may define the quantity $K(x,y)$ analogously as the length of the shortest program which outputs strings x, y, in addition to a means of distinguishing between respective strings. Next to strings, the AIC may be defined for alternative objects: given a computable function f, we define $K(f) = K(i)$, where i is the minimum value i such that the Turing machine T_i computes f. We may interpret this latter definition as the length in bits of the shortest program that computes f.

Note that AIC is uncomputable, which means that in general there exists no algorithm which allows us to obtain the value of $K(\cdot)$, either exactly or as an approximation up to specified precision. As will be discussed in Section 3.2.3, it is however possible to obtain an upper bound on $K(\cdot)$. Furthermore, as discussed in the following Section 3.2.2, we may adopt a probabilistic approach and characterise the expectation value of $K(\cdot)$ with respect to a random variable whose outcomes are binary strings.

3.2.2 Relation to Shannon Information

As previously stated, AIC quantifies information contained in a single object, whereas Shannon information is concerned with the distribution of outcomes in a random variable. This distinction notwithstanding, consider a random variable X whose sample space is the set of finite strings $\mathcal{S} = \{0,1\}^*$. Thus, each outcome $x \in \mathcal{S}$ has an associated probability $P_X(x)$ and an associated

AIC $K(x)$. It may be shown (cf. Grünwald and Vitányi, 2004) that up to specified constants, the expectation $\sum_{x\in\mathcal{S}} P_X(x)K(x)$ approximates the entropy $H(X)$,

$$0 \leq \sum_{x\in\mathcal{S}} P_X(x)K(x) - H(X) \leq K(P_X) + O(1). \tag{3.32}$$

In Equation 3.32, $K(P_X)$ denotes the AIC of the probability mass function P_X, which we assume is computable with respect to outcomes $x \in \mathcal{S}$. As is commonplace in equivalences involving $K(\cdot)$, the constant $O(1)$ depends on the employed reference machine; we may thus interpret Equation 3.32 as stating that for low-complexity distributions, the expectation $\sum_{x\in\mathcal{S}} P_X(x)K(x)$ closely approximates $H(X)$. As discussed by Grünwald and Vitányi (2004), low-complexity distributions are those distributions which are not heavily skewed towards a particular outcome and whose distribution function may be readily described algorithmically.

Owing to Equation 3.32, we might conjecture that for an N-symbol emission from a stationary, finite-order Markov source, the expectation value $\mathbf{E}[X_{1:N}]$ closely approximates its Shannon entropy $H(X_{1:N})$. As demonstrated by Grünwald and Vitányi (2004), it turns out that we have

$$0 \leq \sum_{x\in\{0,1\}^N} P_{X_{1:N}}(x)K(x) - H(X_{1:N}) \leq K(P_{X_{1:N}}) + O(1). \tag{3.33}$$

We may interpret Equation 3.33 as follows: the entropy and expected AIC are equivalent up to the constant $O(1)$ as described in Equation 3.32, plus an additional constant $K(P_{X_{1:N}})$. For a stationary Markov source whose order is much smaller than the considered sequence length N, we treat $K(P_{X_{1:N}})$ as negligible.

3.2.3 The Normalised Information Distance

The normalised information distance (NID) (Li et al., 2004) is a measure of pairwise similarity based on AIC. For finite strings $x, y \in \mathcal{S}$, the NID is defined as

$$\mathrm{NID}(x,y) = \frac{K(x,y) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}}. \tag{3.34}$$

The NID possesses a number of attractive theoretical properties: up to close approximation, it fulfils the metric requirements of identity, symmetry and the triangle inequality (Li et al., 2004). Moreover, the NID characterises dissimilarity using the transformation under which input strings most closely resemble each other. In addition, the NID yields values within the unit interval. Thus, it incorporates the notion that maximally attainable dissimilarity should be invariant of sequence length. Note however that the NID inherits the property of uncomputability from the quantities $K(\cdot)$, $K(\cdot,\cdot)$.

The normalised compression distance (NCD) has been proposed as an approximation of the NID (Li et al., 2004). In the NCD, quantities $K(\cdot)$ are approximated using the number of bits required to represent strings by way of a general-purpose compression algorithm. Similarly, the quantity $K(x,y)$ is approximated by compressing the concatenation of strings x, y.

3.3 Discussion

Having reviewed information-theoretic measures and their properties in the preceding sections, we briefly discuss their use as measures of predictability. Consequently, we justify use of the term *information-theoretic measures of predictability* in the title of this thesis.

Shannon information quantifies predictability in terms of uncertainty about outcomes in a random variable. Given a random variable X, the entropy $H(X)$ quantifies the uncertainty about outcomes in X. Given an information source $\underline{X}$, $H_\mu(\underline{X})$ quantifies the average uncertainty about outcomes in observations emitted by $\underline{X}$, while accounting for correlations among observations. Given two sources $\underline{X}$, $\underline{Y}$, the cross entropy rate $H^\times(\underline{X},\underline{Y})$ and conditional entropy rate $H_\mu(\underline{X}|\underline{Y})$ quantify uncertainty about observations emitted by $\underline{X}$ in relation to observations emitted by $\underline{Y}$, while accounting for correlations among observations. We may interpret the latter measures as quantifying pairwise predictive uncertainty between sources.

The AIC of a string x is the length of the shortest program which outputs x and halts. We may view such a program as the ultimate compressed representation of x. As described informally by Schmidhuber (2009), compression and prediction are closely related: a string which admits efficient compression contains redundant structure; this is in contrast to an algorithmically random string, which possesses no such structure. Efficiently compressing a string amounts to modelling structure which might thereafter be used to form predictions about subsequent, unobserved elements. In a more formally rooted argument, we may appeal to the approximation of expected AIC using Shannon entropy as discussed in Section 3.2.2: for low-complexity distributions, low expected AIC corresponds approximately to low predictive uncertainty; high expected AIC corresponds approximately to high predictive uncertainty. We may interpret the NCD as a measure of pairwise predictability between strings x, y, if we consider the expected AIC of the concatenation of strings xy.

3.4 Conclusion

We have briefly reviewed information-theoretic concepts relevant to our investigations in Chapters 4 and 5. In particular, we have reviewed Shannon's information theory for discrete-valued and continuous-valued random variables, where we consider both sources with and without memory. In addition, we have reviewed the concept of algorithmic information content, whose expected value may be approximated using Shannon information. Furthermore, algorithmic information content may be used to formulate the normalised information distance between two strings.

Chapter 4

Identifying Cover Songs

4.1 Introduction

In Chapter 3, we reviewed measures of predictability and discussed their mathematical properties. Having identified in Chapter 2 information-theoretic measures as a potential means for determining similarity in musical audio, we seek to evaluate such measures for cover song identification. With this aim in view, in this chapter we compare our approach to existing discrete-valued and continuous-valued methods for quantifying predictability between pairs of tracks.

Described in Section 4.2, we propose methods for computing pairwise distances between audio feature sequences, where we consider both discrete-valued and continuous-valued approaches. Our approaches contrast with the normalised compression distance (NCD), a discrete-valued approach which has been evaluated in music content analysis tasks using quantised audio features. In Section 4.2.2, for the discrete case we propose a modification to the NCD, where we account for correlation between sequences. Using artificially generated sequences, we observe that our approach outperforms the NCD as an approximation of the normalised information distance (NID). In Section 4.2.3, for the continuous case we propose to compute information-based measures of similarity as statistics of prediction errors between sequences.

Described in Section 4.3, we evaluate our approaches using a dataset of 300 Jazz standards, in addition to the Million Song Dataset (MSD). For the MSD, to demonstrate scalability we use a *filter-and-refine* approach, based on ranking tracks using a metric distance, and then re-ranking top-ranked result candidates using information-theoretic methods. Thus we obtain results for a

large-scale dataset.

As described in Section 4.4, we observe that continuous-valued approaches outperform discrete-valued approaches. Comparing approaches based on string prediction and compression, we observe that our alignment-based NCD improves performance over existing NCD, for sequential compression algorithms. In addition, we demonstrate that continuous-valued distances may be combined to improve performance with respect to baselines.

4.2 Approach

We denote with $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M)$ two multivariate time series, each representing a sequence of feature vectors extracted from a piece of musical audio. If we assume that both $\mathbf{X}$, $\mathbf{Y}$ consist of independent and identically distributed realisations generated respectively by stochastic processes $\underline{X} = (X_1, X_2, \ldots)$, $\underline{Y} = (Y_1, Y_2, \ldots)$, recall from Chapter 3 that one possible means of quantifying dissimilarity between sequences involves the Kullback-Leibler divergence (KLD), defined as

$$D_{\mathrm{KL}}(p_X \| p_Y) = \int p_X(\mathbf{u}) \log \left(\frac{p_X(\mathbf{u})}{p_Y(\mathbf{u})} \right) d\mathbf{u} \tag{4.1}$$

where $p_X(\mathbf{u})$, $p_Y(\mathbf{u})$ denote the probability density of observation $\mathbf{u}$ emitted by $\underline{X}$, $\underline{Y}$, respectively. Viewed in terms of Shannon information and taking the logarithm to base 2, recall that the KLD quantifies the expected number of additional bits required to represent observations emitted by the memoryless source $\underline{X}$, given an optimal code for observations emitted by memoryless source $\underline{Y}$. As was observed in Chapter 2, the KLD has been widely applied in conjunction with a 'bag-of-features' approach for low-specificity music content analysis tasks (Casey et al., 2008b).

To account for temporal structure in musical audio, as described in Chapter 3 we may use the NCD as a measure of musical dissimilarity between sequences of quantised feature vectors (Li and Sleep, 2005; Ahonen, 2009, 2010; Tabus et al., 2012). Given two strings $x = (x_1, x_2, \ldots, x_N)$, $y = (y_1, y_2, \ldots, y_M)$, the NCD is defined as

$$\mathrm{NCD}(x, y) = \frac{\max\{C(xy) - C(x), C(yx) - C(y)\}}{\max\{C(x), C(y)\}} \tag{4.2}$$

where $C(\cdot)$ denotes the number of bits required to encode a given string, using a compressor such as the Lempel-Ziv (LZ) algorithm (Ziv and Lempel, 1977). Similarly, $C(xy)$ denotes the number of bits required to encode the concatenation of strings x, y. The NCD is an approximation of the

NID (Li et al., 2004), defined as

$$\mathrm{NID}(x,y) = \frac{K(x,y) - \min\{K(x),K(y)\}}{\max\{K(x),K(y)\}} \tag{4.3}$$

where the uncomputable function $K(\cdot)$ denotes algorithmic information content (AIC). The AIC of a given string is the length in bits of the shortest program which outputs the string and then terminates (cf. Li and Vitányi, 2008). Similarly, $K(x,y)$ denotes the length of the shortest program which outputs x, y, in addition to a means of distinguishing between both output strings (cf. Li and Vitányi, 2008). Thus, AIC quantifies the number of bits required to represent specified input strings, under maximally attainable compression.

We are interested in examining the performance of the NCD as an approximation of the NID, where the choice of compressor determines the feature space used to compute similarities in the NCD (Sculley and Brodley, 2006). Furthermore, note that the choice of concatenation in $C(xy)$ to approximate $K(x,y)$ represents an additional heuristic (Li et al., 2004). In the following sections, we describe our contribution.

4.2.1 Quantifying Sequence Dissimilarity Using Shannon Information

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

$$H_\mu(\underline{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_{1:n}) \tag{4.4}$$

$$H_\mu(\underline{X},\underline{Y}) = \lim_{n\to\infty} \frac{1}{n} H((X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)) \tag{4.5}$$

$$= \lim_{n\to\infty} \frac{1}{n} H(X_{1:n},Y_{1:n}) \tag{4.6}$$

$$H_\mu(\underline{X}|\underline{Y}) = H_\mu(\underline{X},\underline{Y}) - H_\mu(\underline{Y}). \tag{4.7}$$

The entropy rate $H_\mu(\underline{X})$ defined in Equation 4.4 quantifies the average amount of uncertainty about X_n, while accounting for correlation between X_n for all n. Analogously, the joint entropy rate $H_\mu(\underline{X},\underline{Y})$ defined in Equation 4.5 quantifies the average amount of uncertainty about the

pair (X_n, Y_n) emitted by sources $\underline{X}, \underline{Y}$, while in addition accounting for correlation between the sources. For the conditional entropy rate $H_\mu(\underline{X}|\underline{Y})$ we have

$$H_\mu(\underline{X}|\underline{Y}) = \lim_{n\to\infty} \frac{1}{n} H(X_{1:n}, Y_{1:n}) - H(Y_{1:n}) \tag{4.8}$$

$$= \lim_{n\to\infty} \frac{1}{n} H(X_{1:n}|Y_{1:n}). \tag{4.9}$$

From Equation 4.9 we may interpret $H_\mu(\underline{X}|\underline{Y})$ as quantifying the average amount of uncertainty about a given emission X_n, while taking into account correlation between observations emitted by $\underline{X}$ and given knowledge of observations emitted by $\underline{Y}$.

For N observations emitted from source X, up to an additive constant the expectation value $\mathbf{E}[K(X_{1:N})]$ may be approximated using the entropy (Grünwald and Vitányi, 2004),

$$\mathbf{E}[K(X_{1:N})] \approx H(X_{1:N}). \tag{4.10}$$

Using Equations 4.4, 4.5 and following Kaltchenko (2004), we assume further approximations

$$\mathbf{E}[K(X_{1:N})] \approx N H_\mu(\underline{X}) \tag{4.11}$$

$$\mathbf{E}[K(X_{1:N}, Y_{1:N})] \approx N H_\mu(\underline{X}, \underline{Y}) \tag{4.12}$$

where $\mathbf{E}[K(X_{1:N}, Y_{1:N})]$ denotes the expected value of $K(\cdot,\cdot)$ for N observations emitted from sources $\underline{X}, \underline{Y}$. In terms of Shannon information, following Kaltchenko (2004) we use Equation 4.7 and estimate the NID as

$$\mathrm{NID}(\underline{X}, \underline{Y}) \approx \frac{\max\{H_\mu(\underline{X}|\underline{Y}), H_\mu(\underline{Y}|\underline{X})\}}{\max\{H_\mu(\underline{X}), H_\mu(\underline{Y})\}}. \tag{4.13}$$

4.2.2 Normalised Compression Distance with Alignment

As given in Equation 4.13, the NID utilises the joint entropy rate $H_\mu(\underline{X}, \underline{Y})$, which accounts for correlation between sources. In contrast, the approach of compressing concatenated strings to estimate $K(x, y)$ may be inadequate for compressors which assume an underlying Markov source, since correlation is not accounted for (Kaltchenko, 2004). To address this possible limitation, we propose the normalised compression distance with alignment (NCDA), defined as

$$\mathrm{NCDA}(x, y) = \frac{C(\langle x, y\rangle) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \tag{4.14}$$

where $\langle a, b\rangle$ performs alignment as a means of maximising correlation between integer-valued strings a, b. Formalised in Algorithm 4.1, we generate equal-length strings by right-padding the

Algorithm 4.1 Pseudo-code for aligning strings a, b. The functions pad() and mode() respectively stand for right-padding and computing the most frequent symbol. See Equation 4.15 for description of $\star$ operator.

if length$(a) >$ length(b) **then**

$b \leftarrow \text{pad}(b, \text{mode}(a), \text{length}(a) - \text{length}(b))$

else

$a \leftarrow \text{pad}(a, \text{mode}(b), \text{length}(b) - \text{length}(a))$

end if

$t \leftarrow \arg\max_\tau((a \star b)[\tau])$

$M \leftarrow \text{length}(a)$

$b_i \leftarrow b_{i+t}$ modulo M, for all $i \in [1..M]$

$\langle a, b \rangle \leftarrow (a_1, b_1, \ldots, a_M, b_M)$

shorter of the two strings with the most common value of the longer string. We then compute the cross-correlation $(a \star b)[t]$, defined as

$$(a \star b)[\tau] = \sum_{m=-\infty}^{\infty} a_m b_{m+\tau} \tag{4.15}$$

where τ denotes the lag, where a_i, b_i respectively denote the ith symbol in a, b and where we define $a_i = 0$, $b_i = 0$ for all $i < 1$. We obtain the value for τ which maximises $(a \star b)[\tau]$, before circularly shifting b by the optimum number of steps. We motivate our choice of cross-correlation by considering that cross-correlation may be computed efficiently, as a series of inner products. Hence, our choice of cross-correlation is pragmatic; an alternative approach might involve estimating mutual information between a and lagged b.

NCDA for Artificial Strings

Before proceeding to an analysis using cover songs as described in Section 4.3, we first examine the behaviour of NCDA and NCD using artificially generated strings.

To facilitate interpretation, we consider the source $\underline{B}$ with equi-probable binary observations $\{0, 1\}$. Since observations are independent, the entropy rate $H_\mu(\underline{B})$ equals the entropy of individual observations $H(B_n)$; since outcomes are equi-probable the entropy is 1 bit. In addition to $\underline{B}$, we consider the source $\underline{B'}$ with each observation b'_n obtained by inverting b_n with probability p. Parameter p thus influences the correlation between $\underline{B}$, $\underline{B'}$. Note that for $p = 0.5$ we obtain minimally correlated $\underline{B}$, $\underline{B'}$.

To obtain an analytical estimate of $\mathrm{NID}(\underline{B},\underline{B'})$, we note that the joint entropy rate $H_\mu(\underline{B'},\underline{B})$ equals the joint entropy of individual observations $H(B_n,B'_n)$, since observations are temporally uncorrelated. Based on the joint probabilities

$$P(B_n=1,B'_n=0)=P(B_n=0,B'_n=1)=p/2 \tag{4.16}$$

$$P(B_n=1,B'_n=1)=P(B_n=0,B'_n=0)=(1-p)/2 \tag{4.17}$$

and using the formula for the joint entropy given in Chapter 3 we obtain for the joint entropy

$$H(B_n,B'_n)=-\left(p\log\frac{p}{2}+q\log\frac{q}{2}\right) \tag{4.18}$$

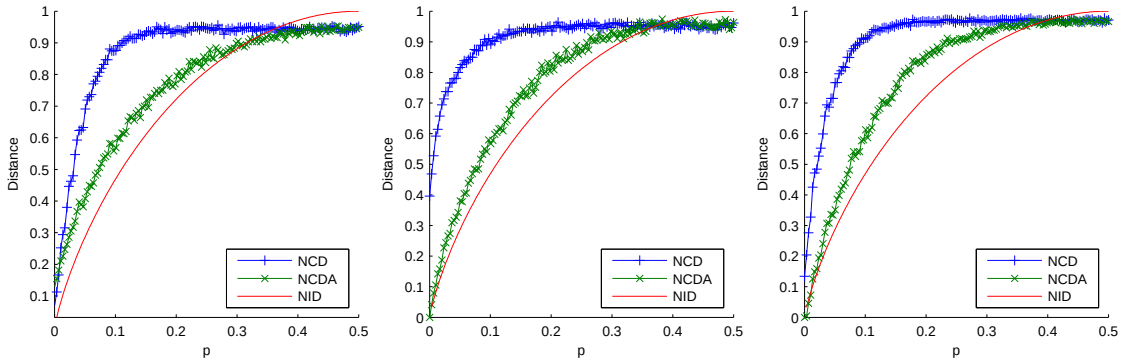
with $q=1-p$. Using Equation 4.13 and the identity $H(X|Y)=H(X,Y)-H(Y)$ we then estimate $\mathrm{NID}(\underline{B},\underline{B'})$ as

$$\mathrm{NID}(\underline{B},\underline{B'})\approx\frac{\max\{H(B_n|B'_n),H(B'_n|B_n)\}}{\max\{H(B_n),H(B'_n)\}} \tag{4.19}$$

$$=H(B_n,B'_n)-1. \tag{4.20}$$

We proceed to examine the behaviour of NCDA and NCD experimentally by generating random binary strings $b_{1:N}$, $b'_{1:N}$ by sampling from the processes B, B' as previously described. We vary p in the range $[0,0.5]$. We compute $\mathrm{NCDA}(b,b')$, $\mathrm{NCD}(b,b')$ using LZ, Burrows-Wheeler (BW) (Burrows and Wheeler, 1994) and prediction by partial matching (PPM) (Cleary and Witten, 1984) compressors, implemented respectively as ZLIB[1], BZIP2[2] and PPMD[3]. As advised by Cebrian et al. (2005), we motivate our choice of string length $N=10^4$ so as not to exceed implementation-defined window sizes used by LZ and BW compressors, which negatively affect compression performance when exceeded. Note that our considered implementation of PPM incorporates methods for modelling conditional symbol probabilities over long contexts (Skibinski and Grabowski, 2004), which are not considered in the original PPM approach proposed by Cleary and Witten (1984). Therefore, we expect no 'pathological' behaviour which would result from computing $\mathrm{NCD}(b,b)$ using a model restricted to short contexts. We employ ZLIB, BZIP2, PPMD with a view to evaluating the performance of NCDA when applied to general-purpose compressors. Our evaluations using compressors are based on representing strings using ASCII encoding. We subsequently compress the obtained textual data.

[1] `http://zlib.org`, retrieved October 2014.
[2] `http://bzip2.org`, retrieved October 2014.
[3] `http://compression.ru/ds/`, retrieved October 2014.



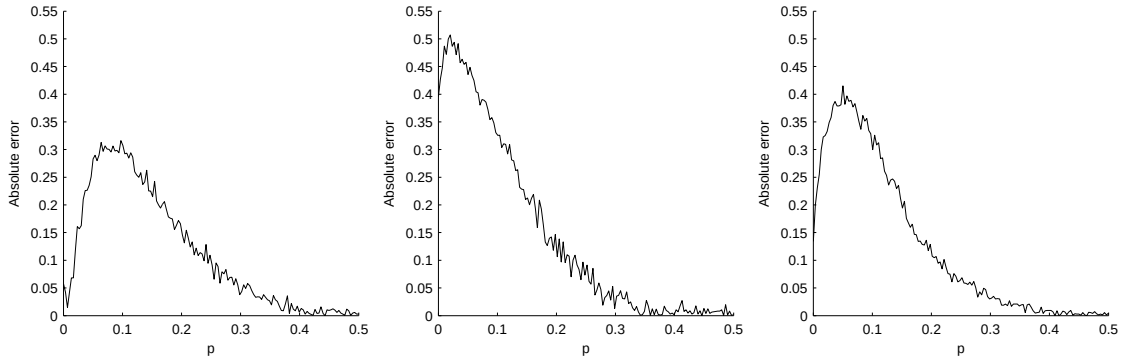
(a) LZ (b) BW (c) PPM

Figure 4.1: Compression distances computed for random strings, with results displayed for LZ, BW, PPM compressors in sub-figures (a)–(c), respectively. Parameter p denotes symbol inversion probability. Each sub-figure includes experimentally obtained NCD, NCDA, in addition to analytically obtained NID.

Figure 4.1 (a)–(c) displays plots of NCD, NCDA in response to p for LZ, BW, PPM algorithms. For comparison, each sub-figure includes the analytically obtained estimate of NID based on Equation 4.19, which we view as our target function. Across compressors, we observe that whereas both NCD and NCDA increase approximately monotonically for increasing p, NCDA more closely approximates NID compared to NCD. Using NCD and for LZ, BW, PPM compressors, we obtain mean absolute absolute errors of 0.17 , 0.20 , 0.19, respectively; using NCDA we obtain corresponding mean absolute errors of 0.07, 0.06, 0.07.

Examining compressors further, we observe for LZ that while NCDA improves performance on average, for small p NCD more closely approximates NID than NCDA. We explain this behaviour by considering that the LZ algorithm identifies repeated substrings (Ziv and Lempel, 1977), a strategy which clearly yields efficient compression for exact repetition in the sequence (b,b). Thus, for near-identical strings we expect NCD to outperform NCDA when used with LZ compression. In contrast, using either PPM or BW and for $p = 0$, NCD incurs target errors of 0.12 and 0.38 compared to jointly 0.00 using NCDA. Since PPM and BW compression do not rely on identifying repeated sequences, this observation matches our expectation of NCDA yielding at least as favourable performance compared to NCD, when dealing with near-identical strings.

Figure 4.2 (a)–(c) displays absolute errors between distances obtained using NCDA and



(a) LZ (b) BW (c) PPM

Figure 4.2: Absolute error between distances obtained using NCDA and NCD with results displayed for LZ, BW, PPM compressors in sub-figures (a)–(c), respectively. Parameter p denotes symbol inversion probability.

NCD. We observe greatest discrepancy between distances for BW compression, followed by PPM and LZ compression. Yet, for BW discrepancy is maximal for comparatively small $p = 0.02$, versus 0.05, 0.09 for PPM, LZ, respectively. Thus, for BW the advantage of using NCDA is for the comparison of near-identical strings. A possible explanation for differences in behaviour between compressors is that our assumptions of a Markov source in NCDA apply less readily to block-based compression schemes such as BW compression, compared to stream-based LZ and PPM.

4.2.3 Predictive Modelling

As previously described, NCD and NCDA rely on determining the number of bits required to encode strings, using a specified compression algorithm. As an alternative approach, we consider the relation between predictability and compressibility (cf. Feder et al., 1992; Feder and Merhav, 1994) and perform sequence prediction. We illustrate our approach for the case of discrete-valued observations. First, recall from Chapter 3 that the entropy rate $H_\mu(\underline{X})$ is defined as

$$H_\mu(\underline{X}) = \lim_{n\to\infty} -\frac{1}{n} \sum_{x_{1:n}\in\mathcal{A}^n} P_{X_{1:n}}(x_{1:n}) \log P_{X_{1:n}}(x_{1:n}) \tag{4.21}$$

where $P_{X_{1:n}}(x_{1:n})$ denotes the probability of observing $X_{1:n} = x_{1:n}$, with $x_{1:n} \in \mathcal{A}^n$ according to the alphabet $\mathcal{A}$. Assume that we have an empirical estimate $\hat{P}_{X_{1:N}}$ of the distribution $P_{X_{1:N}}$, based on a finite number of observations $x_{1:N}$. We estimate $H_\mu(\underline{X})$ using the average log-loss $\ell(\hat{P}_{X_{1:N}}, x_{1:N})$

(cf. Chapter 3), defined as

$$\ell(\hat{P}_{X_{1:N}}, x_{1:N}) = -\frac{1}{N}\log \hat{P}_{X_{1:N}}(x_{1:N}) \tag{4.22}$$

$$= -\frac{1}{N}\left(\log \hat{P}_{X_1}(x_1) + \sum_{i=2}^{N} \log \hat{P}_{X_i|X_{1:i-1}}(x_i|x_{1:i-1})\right) \tag{4.23}$$

where $\hat{P}_{X_i|X_{1:i-1}}(x_i|x_{1:i-1})$ denotes the estimated probability of observing x_i, given preceding context $x_{1:i-1}$. Using Equation 4.23, we thus compute average log-loss by performing a series of predictions based on increasingly long contexts $x_{1:i-1}$. Since $\hat{P}_{X_{1:N}}$ is an estimate of $P_{X_{1:N}}$, the described process is termed *self-prediction* (cf. Serrà et al., 2012).

We denote with $P_{Y_{1:n}}(x_{1:n})$ the probability of observing $x_{1:n}$ from source $\underline{Y}$. Recall from Chapter 3 that a measure of disparity between sources $\underline{X}, \underline{Y}$ is the cross entropy rate $H_\mu^\times(\underline{X}, \underline{Y})$,

$$H_\mu^\times(\underline{X}, \underline{Y}) = \lim_{n\to\infty} -\frac{1}{n} \sum_{x_{1:n}\in\mathcal{A}^n} P_{X_{1:n}}(x_{1:n}) \log P_{Y_{1:n}}(x_{1:n}). \tag{4.24}$$

We estimate $H_\mu^\times(\underline{X}, \underline{Y})$ by computing the average log-loss $\ell(\hat{P}_{Y_{1:N}}, x_{1:N})$ based on iterated prediction, where $\hat{P}_{Y_{1:N}}$ denotes an estimate of $P_{Y_{1:N}}$ based on observations $y_{1:M}$. Since $\hat{P}_{Y_{1:N}}$, $\hat{P}_{X_{1:N}}$ represent disparate sources, the described process is termed *cross-prediction* (cf. Serrà et al., 2012). To obtain a symmetric distance between sources $\underline{X}, \underline{Y}$ based on cross entropy, we compute the quantity

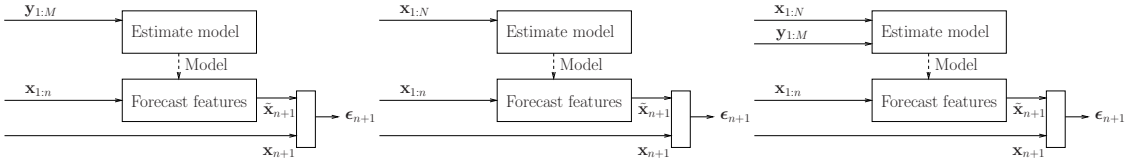
$$D^\times(\underline{X}, \underline{Y}) = \frac{H_\mu^\times(\underline{X}, \underline{Y}) + H_\mu^\times(\underline{Y}, \underline{X})}{H_\mu(\underline{X}) + H_\mu(\underline{Y})} \tag{4.25}$$

where in Equation 4.25 the denominator serves as a normalisation factor, analogous to the denominator in Equation 4.2. We use self-prediction to estimate $H_\mu(\underline{X}), H_\mu(\underline{Y})$.

To obtain a prediction-based estimate of the NID in Equation 4.13, we estimate $H_\mu(\underline{X})$, $H_\mu(\underline{Y})$ again using self-prediction. Furthermore, we estimate the conditional entropy rate $H_\mu(\underline{X}|\underline{Y})$ using the distribution $\hat{P}_{X_{1:N}|Y_{1:M}}$, referring to the estimated distribution of observations $X_{1:N}$ emitted by $\underline{X}$, given knowledge of observations $Y_{1:M} = y_{1:M}$ emitted by $\underline{Y}$. Analogous to self-prediction and cross-prediction, we define the quantity $\ell(\hat{P}_{X_{1:N}|Y_{1:M}}, x_{1:N}, y_{1:M})$,

$$\ell(\hat{P}_{X_{1:N}|Y_{1:M}}, x_{1:N}, y_{1:M}) = \\ -\frac{1}{N}\left(\log \hat{P}_{X_1|Y_{1:M}}(x_1|y_{1:M}) + \sum_{i=2}^{N} \log \hat{P}_{X_i|X_{1:i-1},Y_{1:M}}(x_i|x_{1:i-1}, y_{1:M})\right). \tag{4.26}$$

We refer to the process used to compute Equation 4.26 as *conditional self-prediction*.



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

Figure 4.3: Evaluated prediction strategies. Sequences $\mathbf{x}_{1:N}, \mathbf{y}_{1:M}$ serve as model inputs, observation context $\mathbf{x}_{1:n}$ thereafter forms basis of prediction $\tilde{\mathbf{x}}_{n+1}$. Quantity $\boldsymbol{\epsilon}_{n+1}$ denotes prediction error.

4.2.4 Continuous-Valued Approach

One means of computing the quantities described in Section 4.2.3, involves quantised feature vectors (Li and Sleep, 2005; Ahonen, 2009, 2010; Helén and Virtanen, 2010; Tabus et al., 2012). As an alternative, we propose an approach requiring no prior quantisation.

As used by Serrà et al. (2012), in our approach we utilise non-linear sequence prediction. In contrast to Serrà et al. (2012), we are concerned with evaluating distance measures which we compute as statistics of prediction errors. Therefore, we use a comparatively straightforward nearest-neighbours approach. Given the sequence of feature vectors $\mathbf{C}$, consider first the process of *time delay embedding* (Takens, 1981), which yields the vector sequence $\mathbf{S}^{\mathbf{C}}$, whose elements $\mathbf{s}_r^{\mathbf{C}}$ are defined as

$$\mathbf{s}_r^{\mathbf{C}} = \operatorname{vec}\left(\mathbf{c}_r, \mathbf{c}_{(r-1)\tau}, \ldots, \mathbf{c}_{(r-d+1)\tau}\right). \tag{4.27}$$

According to Equation 4.27, each element $\mathbf{s}_r^{\mathbf{C}}$ aggregates feature vector $\mathbf{c}_r$ along with its preceding temporal context $\left(\mathbf{c}_{(r-1)\tau}, \ldots, \mathbf{c}_{(r-d+1)\tau}\right)$. The amount of temporal context is controlled by parameters d, τ, respectively referred to as *embedding dimension* and *time delay*. The operator vec denotes vectorisation.

Our method of predicting features is based on determining nearest neighbours in time delay embedded space. We first illustrate our method for the case of cross-prediction, depicted schematically in Figure 4.3 (a). Given sequence $\mathbf{y}_{1:M}$, we denote with $\tilde{\mathbf{x}}_{t+h}$ the estimated successor of sequence $\mathbf{x}_{1:t+h-1}$,

$$\tilde{\mathbf{x}}_{t+h} = \mathbf{y}_{q(t)+h} \tag{4.28}$$

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

$q(t)$ as

$$q(t) = \underset{k \in [d\,..\,M-h]}{\arg\max}\ \mathrm{corr}(\mathbf{s}_k^{\mathbf{Y}}, \mathbf{s}_t^{\mathbf{X}}) \tag{4.29}$$

with $\mathrm{corr}(\cdot,\cdot)$ denoting Pearson's correlation coefficient. We motivate use of correlation coefficients as an alternative to the Euclidean distance, following Gómez (2006).

Depicted schematically in Figure 4.3 (b), to perform self-prediction we set $\mathbf{Y} = \mathbf{X}$. Since features may be slowly-varying, when forming prediction $\tilde{\mathbf{x}}_{t+h}$ we disregard observations in the immediate past of time step t. Thus we define

$$\tilde{\mathbf{x}}_{t+h} = \mathbf{x}_{q'(t)+h} \tag{4.30}$$

with $q'(t)$ defined as

$$q'(t) = \underset{k \in [d\,..\,N-h],\, |k-t|>R}{\arg\max}\ \mathrm{corr}(\mathbf{s}_k^{\mathbf{X}}, \mathbf{s}_t^{\mathbf{X}}) \tag{4.31}$$

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

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

$$\tilde{\mathbf{x}}_{t+h} = \mathbf{y}_{q(t)+h}\,\alpha + \mathbf{x}_{q'(t)+h}\,(1-\alpha). \tag{4.32}$$

Similar to the approach given by Foster et al. (2011), in Equation 4.32 we compute the weighting coefficient α using mean squared errors (MSEs),

$$\alpha = \frac{\mathrm{MSE}_{\mathrm{self}}}{\mathrm{MSE}_{\mathrm{self}} + \mathrm{MSE}_{\mathrm{cross}}} \tag{4.33}$$

where $\mathrm{MSE}_{\mathrm{cross}}$, $\mathrm{MSE}_{\mathrm{self}}$ respectively denote cross-prediction and self-prediction mean squared errors. Figure 4.3 (c) depicts conditional self-prediction schematically.

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

$$\varepsilon_{i,n} = \frac{\tilde{x}_{i,n} - x_{i,n}}{\sigma_i^2} \tag{4.34}$$

where σ_i^2 denotes the sample variance of the ith component $(\mathbf{x}_{1:N})_i$ in $\mathbf{x}_{1:N}$. We contrast our approach with the component-wise normalised mean squared error (NMSE) based on cross-prediction used by Serrà et al. (2012), which may be applied as an alternative measure of dissimilarity between sequences. Our approach is based on assuming that the prediction error may be

represented using a normally distributed random variable Z with samples $\boldsymbol{\varepsilon}_{1:N}$. Using the samples, we estimate the continuous entropy $h(Z)$ parametrically. In the case of self-prediction, we assume the approximation $h(Z) \approx H_\mu(\underline{X})$; analogously in the case of cross-prediction and conditional self-prediction, we assume respective approximations $h(Z) \approx H_\mu^\times(\underline{X},\underline{Y})$, $h(Z) \approx H_\mu(\underline{X}|\underline{Y})$. Assuming normality, we estimate $h(Z)$ using the equation

$$h(Z) = \frac{1}{2}\log(2\pi e)^k|\boldsymbol{\Sigma}| \tag{4.35}$$

where $\boldsymbol{\Sigma}$ denotes the sample covariance. In our continuous-valued approach, using the prediction methods depicted in Figure 4.3, we thus estimate information-based measures of uncertainty as statistics of the prediction error sequence. We then substitute the obtained quantities in Equations 4.13 and 4.25 to obtain continuous-valued analogues of the NID and distance $D^\times$.

4.3 Evaluation

We first evaluate our proposed methods using a set of 300 audio recordings of Jazz standards[4]. We assume that two tracks are a cover pair if they possess identical title strings. Thus, we assume a symmetric relation when determining cover identities. The equivalence class of tracks deemed to be covers of one another is a *cover set*. The Jazz data set comprises 97 cover sets, with average cover set size 3.06 tracks.

Furthermore, we perform a large-scale evaluation based on the MSD (Bertin-Mahieux et al., 2011). This dataset includes meta-data and pre-computed audio features for a collection of 10^6 Western popular music recordings. We use a pre-defined evaluation set of 5 236 query tracks partitioned into 1 726 cover sets[5], with average cover set size 3.03 tracks. Following Bertin-Mahieux and Ellis (2012), for each query track, we seek to identify the remaining cover set members contained in the entire 10^6 track collection.

4.3.1 Feature Extraction

For the Jazz dataset, as a representation of musical harmonic content, we extract 12-component beat-synchronous chroma features from audio using the method and implementation described by Ellis and Poliner (2007). We motivate beat-synchronous features as a means of dealing with

[4] `http://www.eecs.qmul.ac.uk/~peterf/jazzdataset.html`, retrieved October 2014.

[5] `http://labrosa.ee.columbia.edu/millionsong/secondhand`, retrieved October 2014.

tempo variation. Assuming an equal-tempered scale, the method accounts for deviations in standard pitch from 440Hz, by shifting the mapping of FFT bins to pitches in the range of ± 0.5 semitones. Following chroma extraction, beat-synchronisation is achieved using the method described by Ellis (2006). First, onset detection is performed by differencing a log-magnitude Mel-frequency spectrogram across time and applying half-wave rectification, before summing across frequency bands. After high-pass filtering the onset signal, a tempo estimate is formed by applying a window function to the autocorrelated onset signal and determining autocorrelation maxima. Varying the centre of the window function allows tempo estimation to incorporate a bias towards a preferred beat rate (PBR). The tempo estimate and onset signal are then used to obtain an optimal set of beat onsets, by using dynamic programming. Chroma features are averaged over beat intervals, before applying square-root compression and normalising chroma features with respect to the Euclidean norm. We evaluate using a PBR of 240 beats per minute (bpm), based on preliminary experiments using NCDA combined with LZ compression. Note that PBR need not relate to the time scale of musical beat, rather we consider PBR the time scale at which we obtain a tempo-invariant representation of chroma features.

The MSD includes 12-component chroma features alongside predicted note and beat onsets (Jehan, 2011), which we use in our evaluations. In contrast to the beat-synchronous features obtained for the Jazz dataset, MSD chroma features are initially aligned to predicted onsets. Based on preliminary evaluations, as an additional processing step we resample predicted beat onsets to match a rate of 240bpm. We then average chroma features over resampled beat intervals. Finally, we normalise features as described for the Jazz dataset.

4.3.2 Key Invariance

To account for musical key variation within cover sets, we transpose chroma sequences using the optimal transposition index (OTI) method (Serrà et al., 2008). Given two chroma vector sequences $\mathbf{X}$, $\mathbf{Y}$, we form summary vectors $\mathbf{h_X}$, $\mathbf{h_Y}$ by averaging over entire sequences. The OTI corresponds to the number of circular shift operations applied to $\mathbf{h_Y}$ which maximises the inner product between $\mathbf{h_X}$ and $\mathbf{h_Y}$,

$$\mathrm{OTI}(\mathbf{h_X}, \mathbf{h_Y}) = \arg\max_i \mathbf{h_X} \cdot \mathrm{circshift}(\mathbf{h_Y}, i) \tag{4.36}$$

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

4.3.3 Quantisation

For discrete-valued similarity measures, we quantise chroma features using the K-means algorithm. We cluster chroma features aggregated across all tracks, where we consider codebook sizes in the range $[2..48]$. To increase stability, we execute the K-means algorithm 20 times. We then select the clustering which minimises the mean squared error between data points and assigned clusters. We observed that the described quantisation method performs similarly to an alternative based on pairwise sequence quantisation.

4.3.4 Distance Measures

We summarise our evaluated distance measures in Table 4.1, where for each distance measure, we list our estimation methods.

As was described in Section 4.2.2, we utilise the following algorithms to compute distance measures by compressing strings: PPM (Cleary and Witten, 1984), BW (Burrows and Wheeler, 1994) and LZ compression (Ziv and Lempel, 1977), implemented respectively as PPMD[6], BZIP2[7] and ZLIB[8]. In all cases, we set parameters to favour compression rates over computation time. To obtain strings for compression, following quantisation we map integer codewords to alphanumeric characters. Subsequently, we represent strings using ASCII encoding, before compressing the obtained data.

We use the described compression algorithms to determine the length in bits of compressed strings and compute NCD, NCDA distances as given in Equations 4.2 and 4.14. In a complementary discrete-valued approach, we use string prediction instead of compression. Using average log-loss, we compute a prediction-based variant of NCDA using the formula

$$\frac{\ell\left(\hat{P}_{\langle X,Y\rangle},\langle x,y\rangle\right)-\min\{\ell\left(\hat{P}_X,x\right),\ell\left(\hat{P}_Y,y\right)\}}{\max\{\ell\left(\hat{P}_X,x\right),\ell\left(\hat{P}_Y,y\right)\}} \tag{4.37}$$

where $\ell\left(\hat{P}_{\langle X,Y\rangle},\langle x,y\rangle\right)$ is the average log-loss obtained from performing self-prediction on the aligned sequence $\langle x,y\rangle$. We compute a prediction-based variant of NCD analogously by predicting concatenated strings without performing any alignment. In addition, we use cross-prediction to estimate distance measure $D^{\times}$, as defined in Equation 4.25. We perform string prediction using the implementations of PPMC and LZ78 algorithms described by Begleiter et al. (2004).

[6] http://compression.ru/ds/
[7] http://bzip2.org
[8] http://zlib.org

Distance	Definition	Estimation method	
NCD	Equation 4.2	String compression (LZ, BW, PPM)	Discrete prediction (LZ, PPM)
NCDA	Equation 4.14	String compression (LZ, BW, PPM)	Discrete prediction (LZ, PPM)
$D^{\times}$	Equation 4.25	Discrete prediction (LZ, PPM)	Continuous prediction
D_{JS}	Equation 4.38	Normalised symbol histograms (*baseline*)	
NID	Equation 4.13	Continuous prediction	
NMSE		Continuous prediction (*baseline*)	
Ellis and Poliner		Continuous cross-correlation (*baseline*)	

Table 4.1: Summary of evaluated distance measures.

Note that the KLD given in Equation 4.1 is non-symmetric. To obtain symmetry, we compute the Jensen-Shannon divergence (JSD) $D_{\mathrm{JS}}(p_X \| p_Y)$, defined as

$$D_{\mathrm{JS}}(p_X \| p_Y) = D_{\mathrm{KL}}(p_X \| p_A) + D_{\mathrm{KL}}(p_Y \| p_A) \tag{4.38}$$

where p_A denotes the mean of p_X, p_Y,

$$p_A = \frac{1}{2}(p_X + p_Y). \tag{4.39}$$

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

We evaluate continuous-valued prediction using parameters $h \in \{1,4\}$, $d \in \{1,2,4\}$, $\tau \in \{1,2,4,6\}$, setting the exclusion radius in Equation 4.31 to $R = 8$ based on preliminary analysis. We compute distance measure $D^{\times}$ using cross-prediction to estimate the numerator in Equation 4.25. In a complementary approach, we estimate the NID using conditional self-prediction to estimate the numerator in Equation 4.13. For $D^{\times}$ and NID, we use self-prediction to estimate the denominator in Equations 4.25, 4.13, respectively.

4.3.5 Performance Statistics

We quantify cover song identification accuracy using mean average precision (MAP), as described by Bello (2011). Relative to a given query track j, we rank all remaining $L-1$ tracks in the data set by ascending distance and denote with $\mathcal{R}(r, j)$ the track at rank r. As an indicator of the relevance of track $\mathcal{R}(r, j)$, we define the binary-valued function $\Omega(r, j)$,

$$\Omega(r, j) = \begin{cases} 1, & \mathcal{R}(r, j) \in \mathcal{C}(j) \\ 0, & \text{otherwise.} \end{cases} \tag{4.40}$$

In addition we define with $\mathrm{P}_j(r)$ the precision at rank r,

$$\mathrm{P}_j(r) = \frac{1}{r} \sum_{c=1}^{r} \Omega(c, j). \tag{4.41}$$

We define with AP_j the average precision for the jth query,

$$\mathrm{AP}_j = \frac{1}{|\mathcal{C}(j)|} \sum_{r=1}^{L} \mathrm{P}_j(r)\Omega(r,j). \tag{4.42}$$

To see why AP_j is useful as a measure of cover song identification accuracy, note that $\mathrm{P}_j(r)$ is the precision (i.e. proportion of covers) that we obtain by considering the top r tracks ranked by ascending distance. We may thus interpret AP_j as the average over all $\mathrm{P}_j(r)$ such that $\mathcal{R}(r,j)$ is a cover. As observed by Aslam et al. (2005), AP_j is an estimate of the area under the precision-recall curve associated with the jth query. By averaging AP_j over all L tracks in the data set as queries, we obtain the MAP.

Following Bello (2011), we use the Friedman test (Friedman, 1937) with Tukey-Kramer post-hoc analysis (Tukey, 1973) to compare accuracies among distance measures. The Friedman test is based on ranking across queries each distance measure according to average precision. We combine the Friedman test with Tukey-Kramer post-hoc analysis to adjust for Type I errors when performing multiple comparisons. We motivate use of both MAP and Friedman test combined with Tukey-Kramer post-hoc analysis to conform to the evaluation procedure used in the MIREX cover song identification task (cf. Downie et al., 2008).

4.3.6 Distance normalisation

To compensate for cover song candidates consistently deemed similar to query tracks, we normalise pairwise distances using the method described by Ravuri and Ellis (2010), based on computing z-scores. For a given distance measure, we denote with $\delta_{i,j}$ the pairwise distance between the ith query track and the jth result candidate. The normalised distance $d_{i,j}$ is obtained as

$$d_{i,j} = \frac{\delta_{i,j} - m_j}{s_j} \tag{4.43}$$

with m_j denoting the average query-wise distance with respect to the jth result candidate,

$$m_j = \frac{1}{L} \sum_{i=1}^{L} \delta_{i,j} \tag{4.44}$$

and with s_j denoting the sample standard deviation of query-wise distances with respect to the jth result candidate,

$$s_j = \sqrt{\frac{1}{L-1} \sum_{i=1}^{L} (\delta_{i,j} - m_j)^2}. \tag{4.45}$$

We apply such distance normalisation as a post-processing step, before computing performance statistics.

4.3.7 Large-Scale Cover Song Identification

For music content analysis involving large datasets, algorithm scalability is an important issue. Since our considered approaches involve non-metric pairwise comparisons between tracks, retrieving result candidates for a given query track requires a linear scan through the dataset, which may be infeasible for large datasets. We use a scalable approach for our evaluations involving the MSD. Following the method proposed by Khadkevich and Omologo (2013), we incorporate our methods into a two-stage retrieval process. By using a metric distance to determine similarity in the first retrieval stage, we allow for the potential use of indexing or hashing schemes, as proposed by Casey et al. (2008a); Schnitzer et al. (2009). We then apply non-metric pairwise comparisons in the second retrieval stage.

In the first stage, we quantise as described in Section 4.3.3 and represent each track with a normalised codeword histogram. Given a query track, we then rank each of the 10^6 candidate tracks using the L1 distance. To account for key variation, for each candidate track we minimise L1 distance across chroma rotations. We then determine the top 1000 candidate tracks, which we re-rank in the second stage using our proposed methods. After both retrieval stages, we normalise pairwise distances as described in Section 4.3.6. We report performance based on the final ranking of all 10^6 candidate tracks, across query tracks.

4.3.8 Combining Distance Measures

To determine whether combining distance measures improves cover song identification accuracy, we obtain pairwise distances as described in Section 4.3.4. We denote with $d^k_{i,j}$ the normalised pairwise distance between the ith query track and the jth result candidate, obtained using the kth distance measure in our evaluation. We transform $d^k_{i,j}$ by computing the inverse rank $d'^k_{i,j}$,

$$d'^k_{i,j} = 1 - \text{rank}(d^k_{i,j})^{-1} \tag{4.46}$$

where $\text{rank}(d^k_{i,j})$ denotes the rank of $d^k_{i,j}$ among all distances obtained with respect to query track i, given the kth distance measure. We apply this transformation to protect against outliers, while ensuring that distance decreases rapidly for track pairs deemed highly similar, for decreasing distance. Note that since our distance transformation preserves monotonicity and MAP itself is based on ranked distances, performance of unmixed distance measures is unaffected by this transformation. Finally, we combine distances $d'^k_{i,j}$, $d'^m_{i,j}$ by computing a weighted average of

distances pooled using max and min operators,

$$\max\{d'^{k}_{i,j}, d'^{m}_{i,j}\}\,\beta + \min\{d'^{k}_{i,j}, d'^{m}_{i,j}\}\,(1-\beta) \tag{4.47}$$

where we optimise weight β with respect to MAP. We subsequently re-normalise the obtained distance, using the method described in Section 4.3.4.

4.3.9 Baseline Approaches

In addition to the JSD and cross-prediction NMSE baselines, we include an evaluation of the method and implementation described by Ellis and Poliner (2007) based on cross-correlation. As a random baseline, we sample pairwise distances from a normal distribution.

4.3.10 Summary

Figure 4.4 summarises our method for cover song identification. Following chroma feature extraction, we obtain key invariance of queries with respect to result candidates, by computing the OTI. Next, we quantise feature sequences using a codebook obtained using the K-means algorithm. We then compute pairwise distances between tracks, where we consider parameter combinations listed in Table 4.1. After normalising obtained pairwise distances, we quantify cover song identification performance using MAP. To evaluate combined distance measures, we use the method described in the preceding Section 4.3.8, before computing performance statistics.

4.4 Results

In Figure 4.5 (a)–(c), based on the Jazz dataset we examine the performance of discrete-valued NCD and NCDA distance measures, combined with LZ, BW and PPM algorithms. For the LZ algorithm, NCDA yields a relative performance gain of 38.6%, averaged across codebook sizes. In contrast, for PPM, with the exception of small codebook sizes in the range $[2..8]$, NCDA yields no consistent improvement over NCD, however averaged across codebook sizes we obtain a mean relative performance gain of 11.0%. Finally, the effect of using NCDA is reversed for BW compression, where performance decreases by an average of 21.8%.

Examining results for the MSD in Figure 4.5 (e)–(g), we observe similar qualitative results for LZ and BW algorithms. For the LZ algorithm, NCDA yields an average relative performance gain of 10.1%, whereas for BW compression we observe an average relative performance loss of

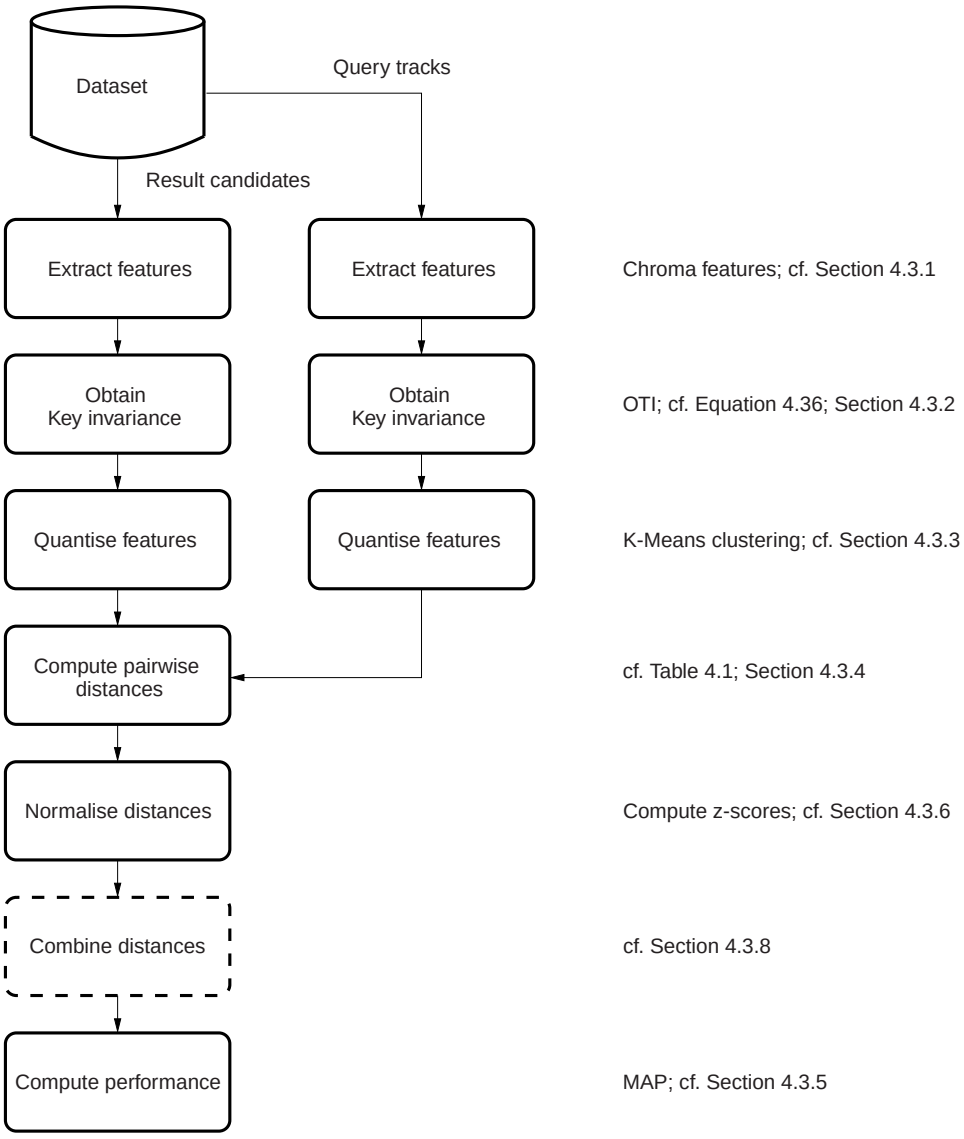


Figure 4.4: Summary of cover song identification method. The dashed box indicates an optional step of combining distances.

6.5%. In contrast to the Jazz dataset, for PPM we observe an average relative performance loss of 1.5%.

For both datasets, compared to NCD, NCDA appears to be advantageous when combined with LZ compression, whereas NCDA combined with BW compression is disadvantageous. Further, NCDA is advantageous combined with PPM, yet only for the case of averaging across codebook sizes and for the Jazz dataset. We note that BW compression is block-based in contrast to LZ and PPM compressors, both of which are stream-based. Following Section 4.2.2, we attribute this property to differences in behaviour among compressors; our assumption of a Markov source in NCDA may apply less readily to BW compression. Concerning differences in relative perfor-

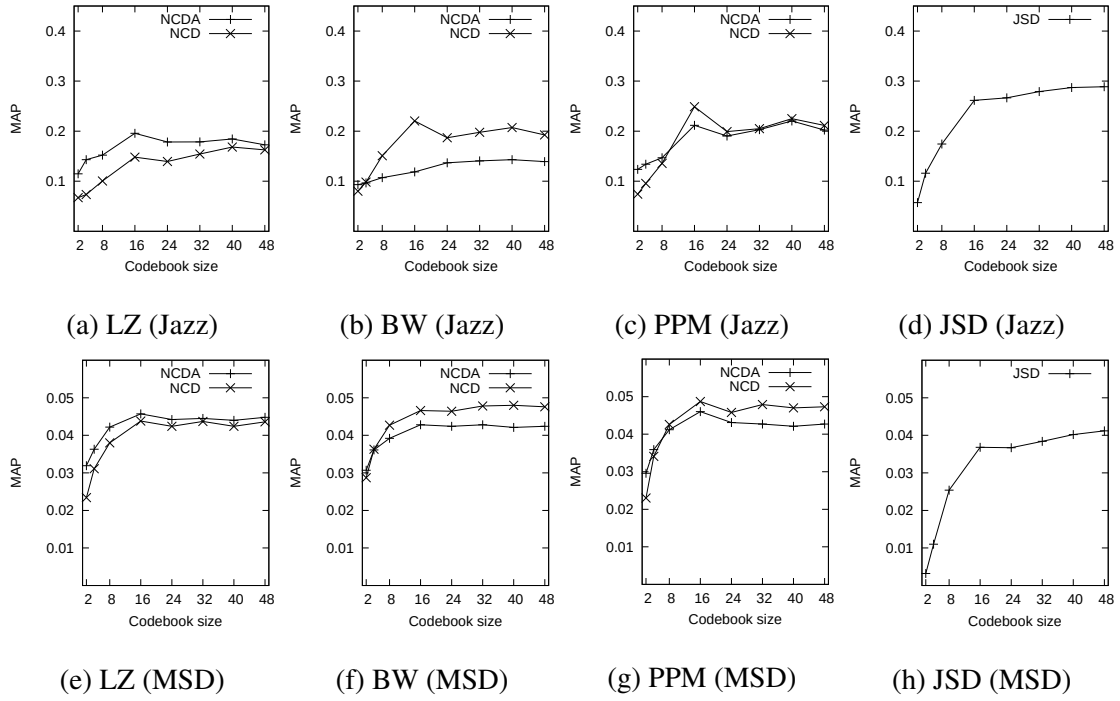
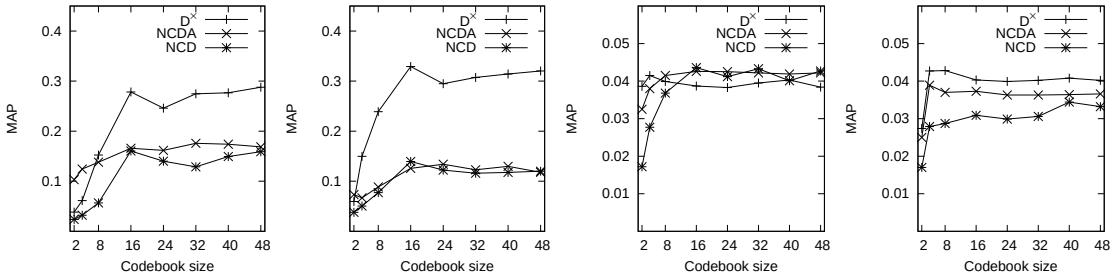


Figure 4.5: Effect of codebook size and distance measure on mean average precision (MAP). Results displayed for LZ, BW and prediction by partial matching (PPM) algorithms in sub-figures (a)–(c), (e)–(g), for Jazz and MSD datasets respectively. Sub-figures (d), (h) display results for Jensen-Shannon divergence baseline (JSD), for Jazz and MSD datasets respectively.

mance gains between datasets, following Khadkevich and Omologo (2013) we further conjecture that chroma feature representation affects the performance of the evaluated distance measures.

We examine the performance of JSD between normalised symbol histograms, as displayed in Figure 4.5 (d), (h). Surprisingly, for the Jazz dataset and for $K > 8$, JSD outperforms compression-based methods, with maximum MAP score 0.289 obtained for $K = 48$. This result is contrary to our expectation that NCD approaches should outperform the bag-of-features approach, by accounting for temporal structure in sequences. In contrast, for the MSD and for optimal K, both NCD and NCDA outperform JSD across all evaluated compression algorithms. We attribute this disparity to differences in problem size between datasets; for the Jazz dataset the problem size may be sufficiently small to amortise advantages of using NCD, NCDA compared to JSD.

In Figure 4.6, we examine the performance of distance measures based on string prediction. For the Jazz dataset, comparing log-loss estimates of NCD and NCDA using the LZ algorithm, averaged across codebook sizes NCDA outperforms NCD; we obtain a mean relative perfor-



(a) LZ (Jazz) (b) PPM (Jazz) (c) LZ (MSD) (d) PPM (MSD)

Figure 4.6: Effect of codebook size and distance measure on mean average precision (MAP). Results obtained using string prediction approach, displayed for Lempel-Ziv (LZ) (sub-figures (a), (c)) and prediction by partial matching (PPM) (sub-figures (b), (d)), for Jazz and MSD datasets respectively.

mance gain of 105.1% (Figure 4.6 (a)). For the PPM algorithm, NCD maximises performance (MAP 0.140); yet in contrast we obtain a mean relative performance gain of 19.3% using NCDA over NCD (Figure 4.6 (b)). Importantly, for both LZ and PPM the cross-prediction distance $D^{\times}$ consistently outperforms NCD and NCDA; for $K = 16$ and combined with PPM compression, we obtain MAP 0.329. For the MSD and using LZ compression, in contrast to the Jazz dataset we observe a mean relative performance loss of 1.8% when comparing $D^{\times}$ with NCDA. For both LZ and PPM, NCDA compared to NCD yields mean relative performance gains of 17.6% and 24.0%, respectively.

Table 4.2 displays the performance of continuous-valued prediction approaches. Note that for $d = 1$, parameter τ may be set to an arbitrary integer following Equation 4.27. We consider results obtained for the Jazz dataset (Table 4.2 (a)–(c)). Using conditional self-prediction to estimate the NID, maximised across parameters h, d, τ we obtain MAP 0.346. In comparison, cross-prediction distance $D^{\times}$ yields MAP 0.454. As a baseline, we determine the cross-prediction NMSE, where maximising across parameters we obtain MAP 0.459. Table 4.2 (a)–(c) displays performance against evaluated parameter combinations. Examining results for the MSD in Table 4.2 (d)–(f), we obtain qualitatively similar results with maximum MAP values 0.0303, 0.0498 and 0.0499 for NID, $D^{\times}$ and NMSE, respectively. For both datasets, we observe that increasing the value of d consistently improves performance. In contrast, we observe no such effect for parameters τ, h.

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

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

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

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

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

(c) NMSE; cross-prediction (Jazz)

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

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

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

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

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

(f) NMSE; cross-prediction (MSD)

Table 4.2: MAP scores for distances based on continuous prediction. In each sub-figure, parameters h, τ, d denote predictive horizon, time delay and embedding dimension, respectively. Results displayed in sub-figures (a)–(c), (d)–(f) for Jazz and MSD datasets, respectively.

4.4.1 Summary of Results and Comparison to State of the Art

Figure 4.7 (a), (b) displays the result of significance testing as described in Section 4.3.5, where we assume 95% confidence intervals and where we maximise across evaluated parameter spaces before testing for significance. Table 4.3 displays a corresponding summary of MAP scores. As baselines we include Ellis and Poliner's cross-correlation approach (Ellis and Poliner, 2007), in addition to randomly sampled pairwise distances.

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

NCD consistently outperforms all discrete-valued approaches for the MSD and significantly outperforms the JSD baseline.

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

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

To facilitate further comparison, we consider the approaches proposed by Bertin-Mahieux and Ellis (2012), Khadkevich and Omologo (2013), who report MAP scores of 0.0295, 0.0371, respectively. Based on such a comparison, we obtain state-of-the-art results. Note that the stated approaches do not report any distance normalisation procedure as described in Section 4.3.4; we found that normalisation improved our results.

Finally, using the method described in Section 4.3.8, we combine distances obtained using continuous-valued prediction. We display results in Table 4.3 and Figure 4.7 (c), (d). Compared to using the baseline cross-prediction NMSE alone, combining NMSE with $D^{\times}$ significantly improves performance for both the Jazz dataset and MSD; we obtain relative MAP performance gains of 8.1% and 3.4% respectively. We obtain no performance gain by further combining NID estimates with NMSE and $D^{\times}$. Using the combination NMSE and $D^{\times}$, we obtain MAP scores 0.496 and 0.0516 for Jazz dataset and MSD, respectively, consistently outperforming the remaining distance measures.

4.5 Conclusion

We have evaluated measures of pairwise predictability between sequences for cover song identification. We consider alternative distance measures to the NCD: we propose NCDA, which incorporates a method for obtaining joint representations of sequences, in addition to methods based on cross-prediction. Secondly, we attend to the issue of representing sequences: we propose continuous-valued prediction as a means of determining pairwise similarity, where we esti-

Dataset	Jazz		MSD	
Method	NCDA	NCD	NCDA	NCD
PPM	0.220	0.249	0.0460	0.0487
BW	0.143	0.220	0.0428	0.0480
LZ	0.196	0.168	0.0457	0.0438
PPM; $D^{\times}$	0.329		0.0428	
LZ; $D^{\times}$	0.288		0.0415	
JSD	0.289		0.0412	
$D^{\times}$ (continuous)	0.454		0.0498	
NID (continuous)	0.346		0.0303	
NMSE (continuous)	0.459		0.0499	
Cross-correlation	0.465		0.0404	
Random	0.026		0.0006	
$D^{\times}$ & NMSE (cont.)	0.496		0.0516	
$D^{\times}$ & NID & NMSE (cont.)	0.432		0.0463	

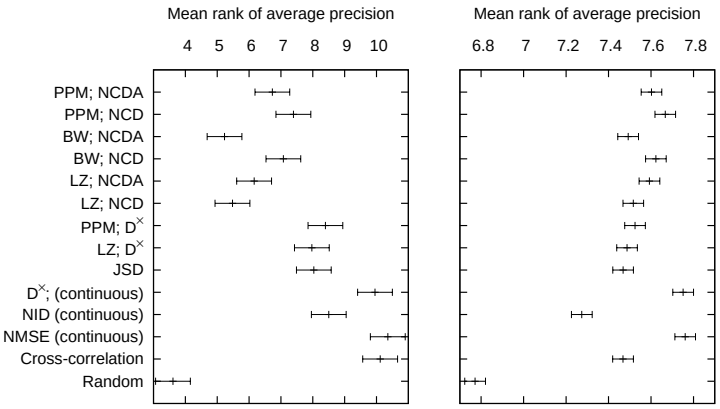
Table 4.3: Summary of MAP scores. First three rows denote compression based approaches. 'Random' denotes sampling pairwise distances from a normal distribution.

mate compressibility as a statistic of the prediction error. We contrast methods requiring feature quantisation, against methods directly applicable to continuous-valued features.

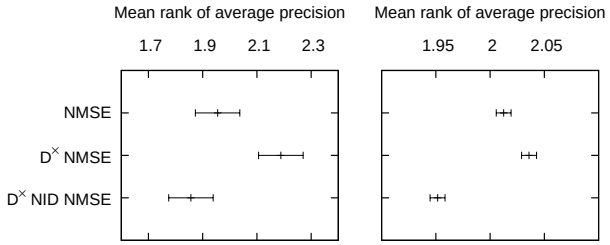
Results indicate that the method of determining pairwise similarity significantly affects cover song identification performance. Firstly, the proposed continuous-valued approach outperforms discrete-valued approaches and competes with evaluated continuous baseline approaches. Secondly, we draw attention to using cross-prediction as an alternative approach to the NCD, where we observe superior results in both discrete and continuous cases for Jazz cover song identification, and for the continuous case for cover song identification using the Million Song Dataset. Thirdly, we have demonstrated state-of-the-art performance using a large-scale dataset. Finally, we have shown that our distances based on continuous-valued prediction may be combined to improve performance relative to the baseline.

Results involving NCDA demand that we distinguish between experiments involving artificial strings and quantised chroma sequences. Whereas we observe performance gains across LZ, BW, PPM algorithms for the case of artificial strings, that NCDA universally improves performance does not hold when considering quantised chroma sequences. For the purpose of cover song identification, in terms of maximally attained performance whereas NCDA is consistently advantageous for LZ compression, NCDA is disadvantageous for BW and PPM compressors. Moreover, we observe a significant reduction in performance for BW, whereas for both LZ and PPM compressors we observe no significant difference in performance using NCDA over NCD.

Contrastingly, in terms of relative performance gains and averaged across codebook sizes, we observe performance gains using NCDA for PPM and for the case of the Jazz dataset. This ob-



(a) Jazz (b) MSD



(c) Jazz (d) MSD

Figure 4.7: Mean ranks of average precision scores obtained using Friedman test. Error bars indicate 95% confidence intervals obtained using Tukey's range test (Tukey, 1973). Higher mean ranks indicate higher performance. Results displayed for Jazz and MSD datasets in sub-figures (a) and (b), respectively, with results for combined distances displayed in sub-figures (c) and (d).

servation holds for approaches based on string compression, as well as string prediction. Further, we observe consistent performance gains using NCDA for PPM and for the case of the MSD, across evaluated codebook sizes. The latter observation holds for approaches based on string prediction. For LZ, in terms of relative performance gains and averaged across codebook sizes NCDA consistently improves performance across evaluated datasets, for approaches based on string compression, as well as string prediction.

We conclude that NCDA may yield performance gains compared to the alternative of NCD, for LZ compression. Considering that we obtain conflicting results for evaluated BW and PPM algorithms, further investigations are however necessary to establish causes for differences between results involving artificial strings and quantised chroma sequences, and for differences between results obtained using string compression versus string prediction. As we suggest in sections 4.2.2 and 4.4, our assumptions of a Markov source may apply less readily to BW com-

pression. We further note that our considered implementations of BW and PPM as general-purpose compressors both incorporate multiple compression steps (cf. Fenwick, 1996; Skibinski and Grabowski, 2004); the influence of individual compression steps should be the subject of future investigations.

Chapter 5

Predicting Musical Similarity

5.1 Introduction

In Chapter 4 we proposed measures of pairwise predictability for cover song identification, a task which we associate with intermediate specificity. In this chapter we estimate predictive uncertainty for the purpose of musical similarity prediction. In particular, we consider the tasks of similarity rating prediction and song year prediction, both tasks which we associate with low specificity.

Described in Section 5.2, our approach is based on computing track-wise descriptors on audio feature sequences. Specifically, we propose to use measures of predictability as track-wise statistics of audio feature sequences. Thus, this approach contrasts with Chapter 4, where we use predictive uncertainty to quantify pairwise similarity between feature sequences. In this chapter, we take the obtained track-wise measures of temporal regularity as our feature space for classification and regression, for our chosen tasks.

Section 5.3 details our evaluations. We describe our method and results for similarity rating prediction in Section 5.3.1; we describe our method and results for similarity rating prediction and song year prediction in Section 5.3.2. For both considered tasks, we observe that our descriptors capture musically relevant information and that our descriptors improve predictive accuracy with respect to baseline approaches.

5.2 Approach

Assume that we have the audio feature vector sequence $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, computed for a given piece of music. As was discussed in Chapter 2, as a feature descriptor we may compute statistical moments of $\mathbf{X}$, yet such bag-of-features representation disregards the temporal order of feature vectors. As our descriptor for predicting musical similarity, we propose the compression rate $R_\lambda(\mathbf{X})$,

$$R_\lambda(\mathbf{X}) = \frac{C(Q(\mathbf{X}, \lambda))}{T} \tag{5.1}$$

where $C(\cdot)$ denotes the number of bits required to represent a string and where we quantise $\mathbf{X}$ using the quantisation scheme Q with λ levels. Note from Chapter 3 that we may take $C(\cdot)$ as an estimate of the algorithmic information in $Q(\mathbf{X}, \lambda)$; moreover we may take $R_\lambda(\mathbf{X})$ as an estimate of the entropy rate of $Q(\mathbf{X}, \lambda)$. To control for variation in sequence length in our evaluations, rather than simply use $C(\cdot)$ we normalise with respect to T. Finally, we favour quantisation over continuous-valued measures of predictability as described in Chapter 4, using standard time and memory-efficient string compressors when computing $C(\cdot)$ in our evaluations. Thus, we motivate use of discrete-valued approaches on pragmatic grounds in our investigations; we view using continuous-valued prediction as an alternative approach which should be explored in future work.

Given a 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 with 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 the downsampling factors $1, 2, 4, 8$. We distinguish among temporal resolutions using the labels FCD1, FCD2, FCD4, FCD8, respectively. For features with variable frame rate, we compute FCDs with no further downsampling applied. Algorithm 5.1 lists pseudo-code for computing FCDs with respect to downsampling factor F, feature sequence $\mathbf{X}$, quantisation granularity λ.

As 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 (Figure 5.1 (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 (Figure 5.1 (c)), we obtain high values for R_λ,

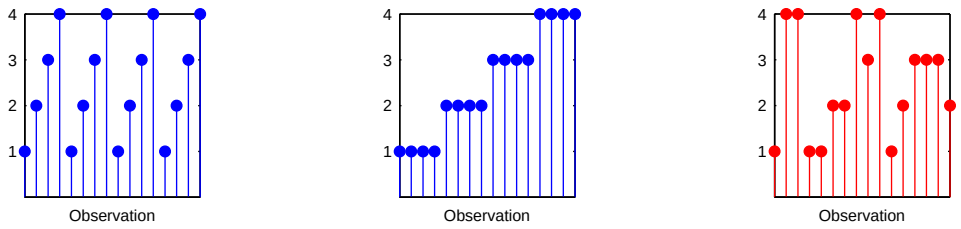
Algorithm 5.1 Pseudo-code for computing FCDs with respect to downsampling factor F, feature vector sequence $\mathbf{X}$, λ quantisation levels. Quantity $C(s)$ denotes number of bits required to represent string s using a given string compressor. See main text for description of quantisation function $Q(\cdot,\cdot)$.

$\mathbf{X} \leftarrow \text{downsample}(\mathbf{X}, F)$

string s $\leftarrow Q(\mathbf{X}, \lambda)$

FCD $\leftarrow C(s)/\text{length}(s)$

since the quantised feature sequence no longer admits an efficient encoding. We thus consider FCDs a statistic quantifying the amount of temporal regularity in a feature sequence. Note from Figure 5.1 that FCDs are invariant to any re-scaling of feature values. In contrast, statistical moments such as mean and variance are invariant to any re-ordering of features. We observed in Chapter 2 that feature moments have been widely applied for low-specificity 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 combined with feature moment descriptors (FMDs) may be used to improve prediction accuracy with respect to using feature moments alone, for our considered tasks.



(a) Low R_λ (b) Low R_λ (c) High R_λ

Figure 5.1: Hypothetical sequences with low and high R_λ, assuming $\lambda = 4$.

5.2.1 Similarity Rating Prediction

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

for the ith 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

$$P\left(S = k | \mathbf{d}_{\langle i,j \rangle}\right) = \frac{\exp\left(\boldsymbol{\beta}_k^T \mathbf{d}_{\langle i,j \rangle} + \gamma_k\right)}{\sum_{m=1}^{K} \exp\left(\boldsymbol{\beta}_m^T \mathbf{d}_{\langle i,j \rangle} + \gamma_m\right)} \tag{5.2}$$

where $\boldsymbol{\beta}_k$, 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 $P\left(S = k | \mathbf{d}_{\langle i,j \rangle}\right)$. In this way, we retain the ordinal scale of training samples, when forming predictions. We describe our model estimation method in Section 5.3.1.

5.2.2 Song Year Prediction

For the task of song year prediction, we hypothesise that descriptor values correlate with the chart entry dates of tracks. Following Bertin-Mahieux et al. (2011) we apply a linear regression model. Given the ith track in our collection, we predict the associated chart entry date y_i using a linear combination of components in descriptor vectors $\mathbf{r}_{i,n}$,

$$\hat{y}_i = \sum_{n=1}^{N} \boldsymbol{\theta}_n^T \mathbf{r}_{i,n} + \alpha \tag{5.3}$$

where $\boldsymbol{\theta}_n$ denotes regression coefficients for the nth 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 5.3.2. 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.

5.3 Evaluation

For our evaluations, we use a collection of 15 473 entries from the American *Billboard Hot 100* singles popularity chart[1]. Each entry in the dataset is represented by a track excerpt of

[1] `http://www.billboard.com`, retrieved October 2014.

Feature name	Description
Chroma (Ellis and Poliner)	12-component chromagram based on using phase-derivatives to identify tonal components in spectrum (Ellis and Poliner, 2007).
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 (Plomp and Levelt, 1965) between peak pairs in magnitude spectrum.
spectral.irregularity	Squared amplitude difference between successive partials (Jensen, 1999).
spectral.mfcc	12-component Mel-frequency cepstral coefficients (MFCCs) (Slaney, 1998) (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 (Masri, 1996).
tonal.chromagram.centroid	Centroid of 12-component chromagram.
tonal.keyclarity	Peak correlation of chromagram with key profiles (Gómez, 2006).
tonal.mode	Predicted mode after correlating chromagram with key profiles.
tonal.hcdf	Flux of 6-dimensional tonal centroid (Harte et al., 2006).

Table 5.1: Summary of evaluated audio features.

approximately 30s of audio, and is annotated with a chart entry date. Chart entry dates span the years 1957–2010 ($M = 1982.9$y, $SD = 15.4$y). Our choice of the Billboard dataset is motivated by prior availability of web-sourced similarity ratings for a subset of 7 784 track excerpts.

For each track excerpt in the dataset, we extract a set of 25 audio features, using MIRToolbox (Lartillot and Toiviainen, 2007) version 1.3.2 and using the framewise chromagram representation proposed by Ellis and Poliner (2007). With the exception of rhythmic features, which are computed using predicted onsets, features are based on a constant frame rate of 40Hz. Table 5.1 summarises the set 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 FMDs. We compute FCDs as described in Section 5.2, where for the

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 5.2: Artists ranked by 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.

case of the vector-valued features chroma, MFCCs and delta-MFCCs we apply principal components analysis (PCA) in track-wise fashion as an additional 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 matching (PPM) algorithm[2], based on the implementation described in Begleiter et al. (2004). 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 (LZ) compression (Ziv and Lempel, 1978). Nevertheless, we note that PPM compresses efficiently compared to alternative compression schemes (Begleiter et al., 2004).

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 5.2, we report the 20 lowest-ranking and highest-ranking artists. Additionally, across artists we report tracks with median FCD scores ('medoid tracks').

Comparing track groups, the lowest-ranking artists are predominantly vocalists with a repertoire of Jazz ballads and slow-moving pieces, with smooth timbral characteristics (e.g. Johnny Mathis, Barbara Streisand). In contrast, the artists with highest complexity values stand for music with strong percussive and aggressive components, from surf-rock (Jan & Dean), through

[2] `http://www.cs.technion.ac.il/~ronbeg/vmm/index.html`, retrieved October 2014.

1980s Power Rock (Van Halen) and Hip Hop (Eminem). Our own 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 our considered similarity tasks. Note however that we make no claim that FCDs capture any notion of musical complexity as discussed by Pressing (1999). While beyond the scope of our work, track-wise analysis of FCDs merits further investigation.

5.3.1 Similarity Rating Prediction

We evaluate similarity rating prediction using annotations collected for a subset of the chart music dataset. For our investigations, we use an existing collection of 7 784 pairwise similarity ratings from 456 subjects participating in a web-based listening test[3]. 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 qualifying 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 (Wiggins et al., 2010), human similarity judgements are context-dependent (Goodman, 1972; Tversky, 1977). 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,

[3] `http://webprojects.eecs.qmul.ac.uk/matthiasm/audioquality-pre/check.php`, retrieved October 2014.

which might affect similarity ratings (Sturm, 2012). We obtain a median of 6 ratings per subject, with each rating corresponding to a unique track pair. Table 5.3 displays obtained score counts.

As shown in Table 5.3, 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 by Bogdanov et al. (2011), 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 (Downie et al., 2010). 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.

	Similarity score				
	1	**2**	**3**	**4**	**5**
Count	2 060	2 115	1 742	1 391	476

Table 5.3: 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) Ollen (2006). We obtain a median OMSI score 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 prior ratings. Across subjects, we obtain a median of 42 ratings ($M = 45.4$, $SD = 29.3$), after discarding ratings for duplicated track pairs. We aggregate controlled-condition ratings across participants and again discard ratings for duplicated track pairs. Across subjects we thus obtain a total of 509 controlled-condition similarity ratings, corresponding to 6.5% coverage of web-based similarity ratings. Table 5.4 displays a 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

		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 5.4: Confusion matrix of web-sourced versus controlled-condition similarity ratings.

we report results for the four-point rating scale in brackets. We first quantify agreement using Kendall's correlation coefficient τ_b (cf. Agresti, 2010). 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 (cf. Efron, 1982). 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 (cf. Agresti, 2010), 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 5.4 and interpreting the controlled-condition rating process as a multinomial classification task, we obtain a balanced accuracy (BA) of 0.292 (0.345); the corresponding 95% confidence interval is $[0.254, 0.336]$ ($[0.304, 0.393]$).

Distance Measures

We predict similarity ratings by applying multinomial regression to pairwise Euclidean distances between descriptor vectors, using the approach described in Section 5.2.1. As an additional baseline distance measure, assuming Gaussianity and diagonal covariance, we compute the Kullback-Leibler divergence (KLD) on pairs of FMDs. Given means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and variances $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$, each with dimensionality c, we compute the KLD in closed form as

$$\mathrm{KLD} = \frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\right) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - c - \log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right). \tag{5.4}$$

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. As described in Chapter 4, following Serrà et al. (2012), we apply time delay embedding (Takens, 1981) separately to pairs of feature sequences. We restrict our evaluated parameter space by setting unit time delay in Equation 4.27. Given feature vectors $(\mathbf{v}_1, \ldots, \mathbf{v}_T)$ each with dimensionality c, time delay embedding then produces higher-dimensional feature vectors with dimensionality dc by stacking d consecutive vectors $(\mathbf{v}_{t-d}, \ldots, \mathbf{v}_{t-1})$ at each time step t. We perform cross-predictions by determining sequential successors of nearest neighbours in the embedded space, using the approach described in Equations 4.28, 4.29. We restrict our evaluated parameter space by setting unit predictive horizon in Equation 4.29. As our distance measure between predicted and observed feature sequences, we compute the normalised mean squared error, as described in Section 4.2.4. 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.

Performance Statistics

To quantify the accuracy of similarity rating prediction, as discussed by Cardoso and Sousa (2011) we compute Kendall's τ_b and Spearman's ρ_s, both which are ordinal measures of association between predicted and annotated similarity ratings. As given by Agresti (2010), we define Kendall's τ_b as follows: assume that we have sequences $\mathcal{Q} = (q_1, \ldots, q_M)$, $\mathcal{O} = (o_1, \ldots, o_M)$. 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

$$\tau_b = \frac{M_c - M_d}{\sqrt{(M_p - M_q)(M_p - M_o)}} \tag{5.5}$$

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 $\mathcal{Q}$, $\mathcal{O}$ (Agresti, 2010). 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 (Pinto da Costa et al., 2008). 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 our notion of relative rating agreement, we thus consider BA a subsidiary measure of performance compared to τ_b, ρ_s.

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 5.5. We standardise distances with respect to 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 5.5, we obtain 150 distances. As given in Equation 5.2, we weight distances individually.

In our training step, we estimate multinomial regression parameters using elastic net regularisation (ENR) (Zou and Hastie, 2005) based on coordinate descent (Friedman et al., 2010; Qian et al., 2013). We denote with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^T$ regression coefficients and model intercepts as given in Equation 5.2. Using ENR, we solve

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \left\{ \eta \left(\nu \|\boldsymbol{\beta}\|_1 + (1-\nu)\frac{1}{2}\|\boldsymbol{\beta}\|_2^2 \right) - \ell(\boldsymbol{\beta},\boldsymbol{\gamma}) \right\} \tag{5.6}$$

where $\ell(\boldsymbol{\beta},\boldsymbol{\gamma})$ 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 τ_b, ρ_s, BA and for each rating scale, we apply hold-out

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 5.5: 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).

validation to training data and optimise η by determining maximal prediction accuracy. We consider ν a hyper-parameter which we assign constant value 0.444 after a single optimisation step; we optimise Kendall's τ_b with respect to the five-point rating scale and using a model incorporating FMDs, where we again apply hold-out validation to training data.

Summary

Figure 5.2 summarises our method for similarity rating prediction. Following feature extraction (cf. Table 5.1), we compute FCDs and FMDs. Next, we compute pairwise distances between FCDs and FMDs. Using ENR, we estimate a multinomial regression model of similarity ratings in response to pairwise distances; we consider combinations of descriptors and distance measures as given in Table 5.5. Using cross-validation, we consider as performance measures ρ_s, τ_b, BA.

Results

Figure 5.3 displays the result of exploratory analysis, in which we plot pairwise distances against similarity ratings. We consider FCDs computed without downsampling and FMDs, respectively compared using Euclidean distance and log-transformed KLD. For both descriptors, we average distances across across features, to identify any possible relationship between similarity rating and pairwise distances between descriptors. We observe a monotonically decreasing trend in median, 25th and 75th percentile ranges against increasing similarity rating, suggesting that both FCDs and FMDs may be used to predict similarity ratings.

We examine the correlation between descriptor distances and five-point scale similarity ratings across individual audio features. Figure 5.4 depicts correlations τ_b for FCDs and FMDs, where we compare FMDs using both Euclidean distance and KLD. In addition to FMDs, as described we consider as a baseline the cross-prediction error.

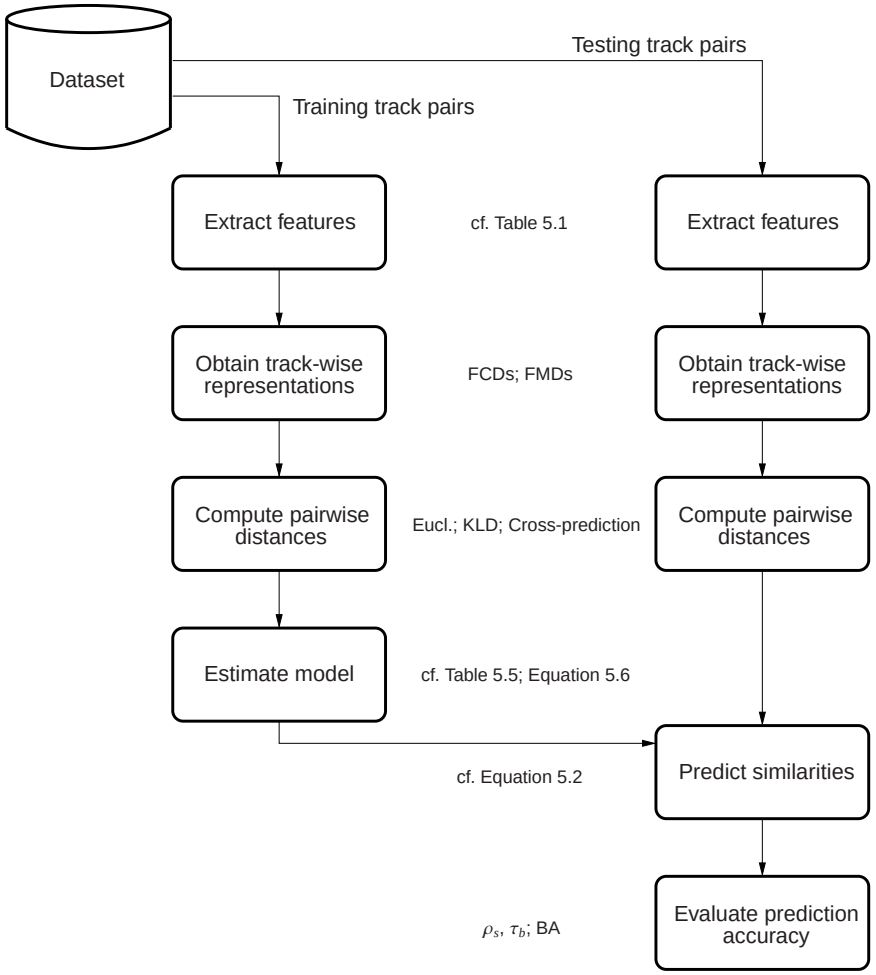
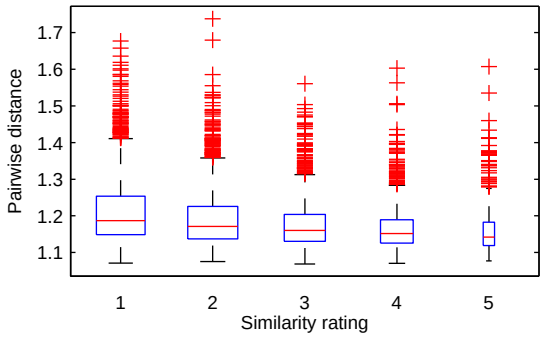


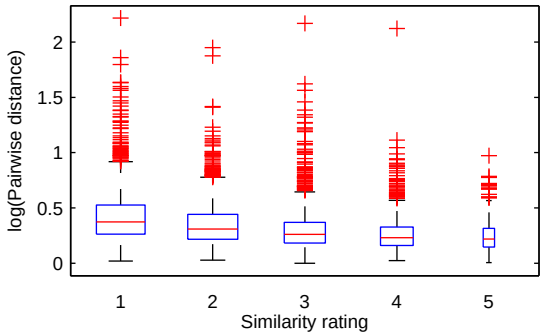
Figure 5.2: Summary of similarity rating prediction method.

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.

Figure 5.5 displays a comparison of similarity rating prediction accuracy, where for each descriptor set in Table 5.5 we apply feature selection using ENR. We estimate models using



(a) FCDs (FCD1) compared using Euclidean distance



(b) FMDs compared using KLD

Figure 5.3: Box plot of pairwise distances against web-sourced pairwise similarity ratings, obtained using (a) FCDs computed without downsampling and (b) FMDs. Distances averaged across features. Crosses represent outliers. Box widths proportional to number of observations.

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 rating scales, we observe that FCDs are outperformed by FMDs compared using KLD alone, or FMDs compared using Euclidean distance and KLD in combination: Compared to FCDs and based on the five-point rating scale, for FMDs compared using aggregated Euclidean distance and KLD we observe absolute performance gains of 0.018, 0.043, 0.039 with respect to ρ_s, τ_b, BA; the respective relative performance gains are 7.0%, 19.3%, 14.1%. Comparing analogously for the four-point rating scale, using FMDs in place of FCDs we observe absolute performance gains of 0.009, 0.016, 0.017; the respective relative performance gains are 3.2%, 10.1%, 9.8%.

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

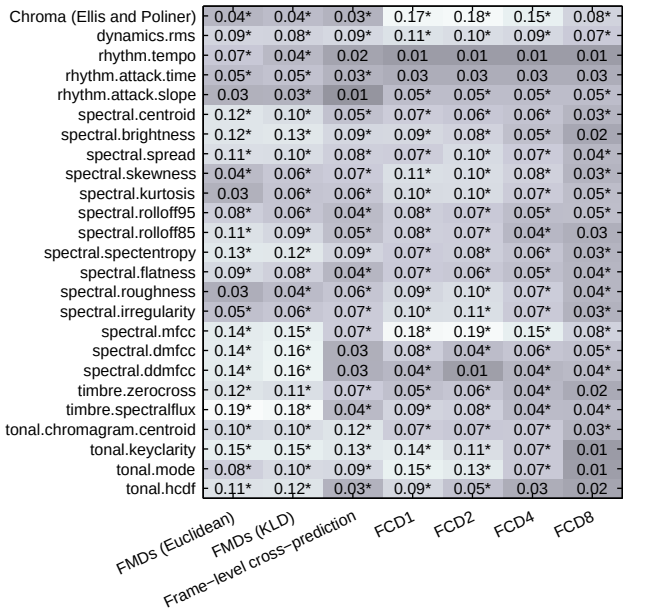
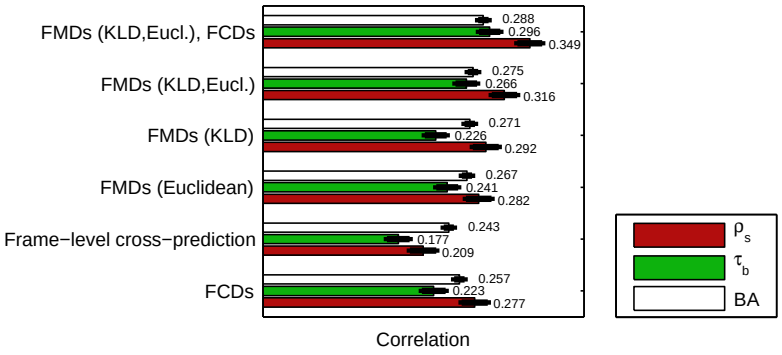


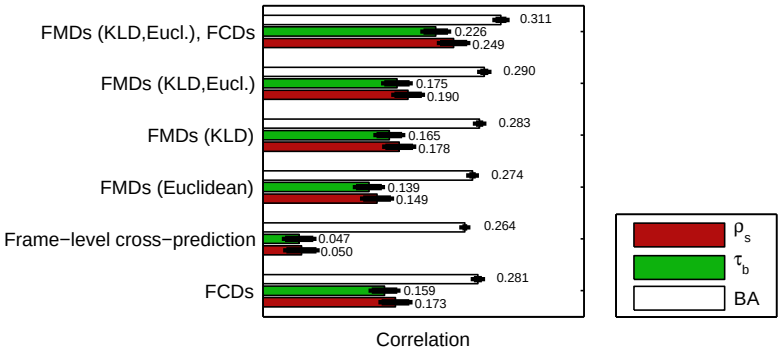
Figure 5.4: Feature-wise absolute correlation $|\tau_b|$ between pairwise distances and web-sourced similarity annotations. Pairwise distances respectively obtained using FMDs compared using Euclidean distance and KLD (first and second columns), cross-prediction (third column), Euclidean distance applied to FCDs (remaining columns). Starred entries indicate significance, where we apply Bonferroni correction to $\alpha = 0.05$.

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%. As suggested by Figure 5.4, we observe that cross-prediction yields comparatively low prediction performance. 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 (ANOVA) 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. We use ANOVA instead of analogous non-parametric tests as applied in Chapter 4, since visual inspection of histograms revealed that bootstrapped statistics are normally distributed with approximately equal variance, thus fulfilling ANOVA's assumptions.

Figure 5.6 displays regression coefficients across features and descriptor classes, where we consider the best-performing model evaluated in Figure 5.5 based on ρ_s and using the five-point rating scale. We sum regression coefficient magnitudes across each of the K binary classifiers



(a) Five-point rating scale



(b) Four-point rating scale

Figure 5.5: Similarity rating prediction accuracy using combined descriptors. Standard errors obtained by bootstrap sampling pairs of predicted and observed similarity ratings.

given in Equation 5.2, before normalising the obtained values to sum to one. Comparing 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.

5.3.2 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 Equation 5.3. Figure 5.7 displays a histogram of chart entry dates.

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

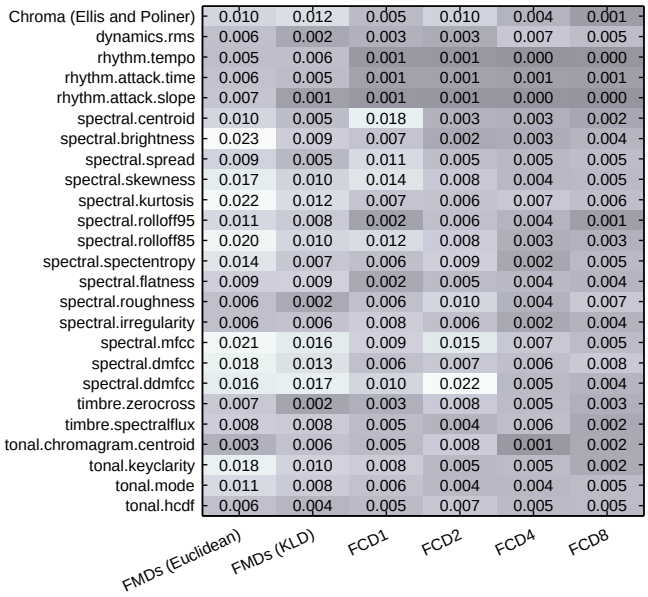


Figure 5.6: Normalised regression coefficient magnitudes, estimated using elastic net regression, for task of similarity rating prediction. Candidate descriptor set comprised of FCDs compared using Euclidean distance, and FMDs compared using Euclidean distance and KLD.

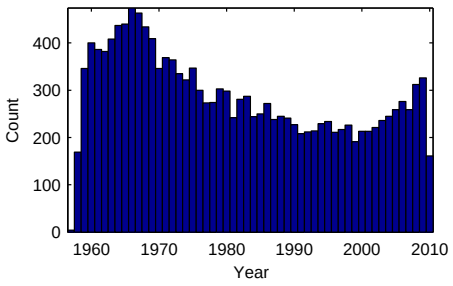


Figure 5.7: Histogram of chart entry dates.

effects, in addition to any analogous effects at the level of artists Flexer and Schnitzer (2010). The resulting training and testing datasets consist of 10 728 and 4 745 tracks respectively. We deem as outliers descriptor values in the training data exceeding 10 standard deviations beyond the 99th percentile. We replace such outliers with imputed values, using the K-nearest neighbours (KNN) algorithm.

We apply linear regression separately to sets of descriptor vectors, as specified in Table 5.6. We standardise descriptors with respect to 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

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 5.6: Summary of descriptor combinations evaluated for song year prediction. Fourth column lists number of coefficients in linear regression model (excluding intercept).

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 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_N^T)^T$, α regression coefficients and the model intercept as given in Equation 5.3. Using ENR, we solve

$$\min_{\boldsymbol{\theta},\alpha} \left\{ \eta \left(\nu \|\boldsymbol{\theta}\|_1 + (1-\nu)\frac{1}{2}\|\boldsymbol{\theta}\|_2^2 \right) + \mathrm{SSR}(\boldsymbol{\theta},\alpha) \right\} \tag{5.7}$$

where $\mathrm{SSR}(\boldsymbol{\theta},\alpha)$ denotes the sum of squared residuals. Both η, ν behave as defined in Equation 5.6. We apply cross-validation to training data and optimise η by determining minimal prediction mean squared error. We again consider ν a hyper-parameter which we assign constant value 0.163; we optimise prediction mean squared 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 $[1957\mathrm{y}\,..\,2010\mathrm{y}]$.

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, 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

$$\mathbf{r}'_{w,n} = \frac{1}{|C_w|} \sum_{i \in C_w} \mathbf{r}_{i,n} \tag{5.8}$$

where C_w denotes the set of tracks at window position w. We apply the windowing procedure separately to training and testing data sets. For a given window size, we estimate our model using ENR as previously described; given the obtained regression model and given descriptor vectors at window position z in the testing data, we seek to predict the associated window centre y'_z.

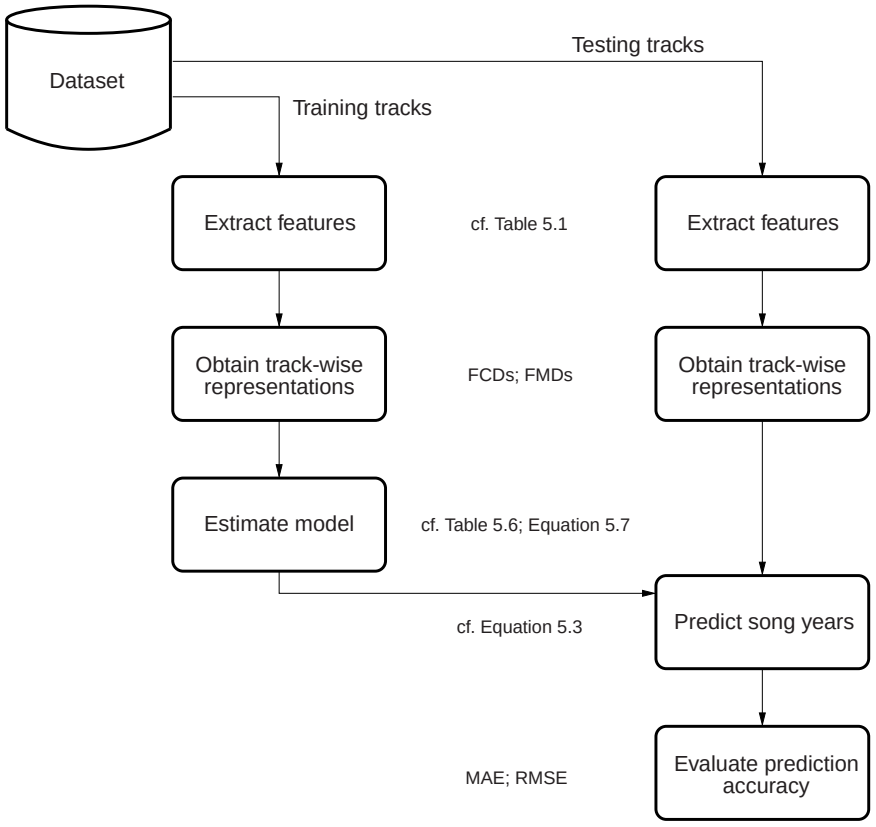


Figure 5.8: Summary of song year prediction method.

Performance Statistics

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

Summary

Figure 5.8 summarises our method for song year prediction. Following feature extraction (cf. Table 5.1), we compute FCDs and FMDs. Next, using ENR, we estimate a linear regression model of chart entry dates in response to feature values; we consider combinations of descriptors and distance measures as given in Table 5.6. Using cross-validation, we consider as performance measures ρ_s, τ_b, BA.

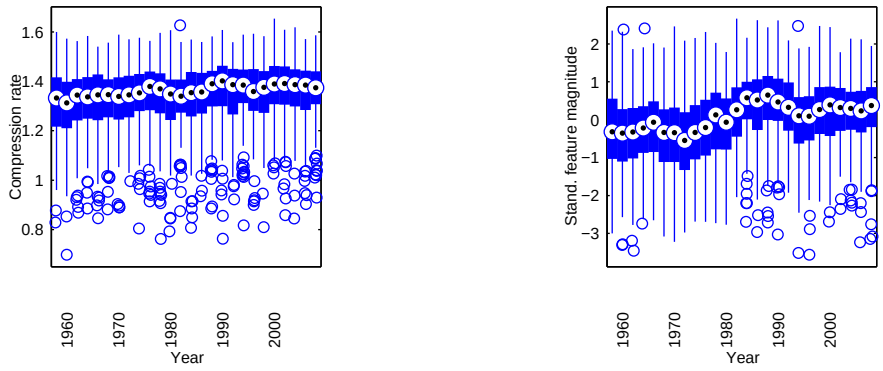
Results

Figure 5.9 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 (cf. Lartillot and Toiviainen, 2007). The resulting year-wise box plots suggest that the examined descriptors

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 5.7: 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.

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 sequences for lags in the range $[1\,..\,15]$, we observe weaker correlations for FCDs compared to FMDs. Yet, both autocorrelations exhibit slowly decaying autocorrelations (Figure 5.10), characteristic of a non-stationary sequences (cf. Kirchgassner et al., 2012). Following the method of Box and Jenkins (cf. Box et al., 2013), we attempt to attain stationarity by applying first-order differencing to the sequences. However, we observe autocorrelation close to -0.5 at unit lag, suggesting that the sequences have been overdifferenced (cf. Kirchgassner et al., 2012). We interpret these observations as evidence for a non-trivial, trend-exhibiting process governing observed descriptor values (Granger and Joyeux, 1980).

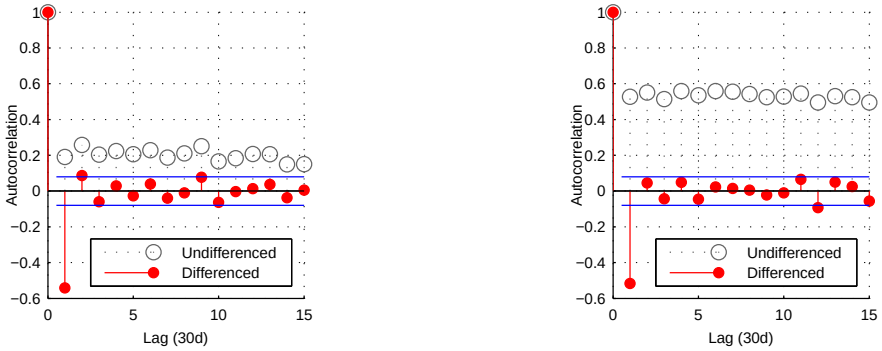


(a) FCDs (FCD1)

(b) FMDs

Figure 5.9: 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.

Table 5.7 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



(a) FCDs (FCD1)

(b) FMDs

Figure 5.10: 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 sequences.

prediction using FCDs alone: We obtain respective performance gains of 12.3% (MAE) and 9.4% (RMSE). 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 5.3.1, 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 ANOVA 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. As previously described in Section 5.3.1, we motivate ANOVA on the basis that exploratory analysis of bootstrapped statistics revealed normality and approximately equal variance.

Figure 5.11 displays regression coefficients obtained using unwindowed 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

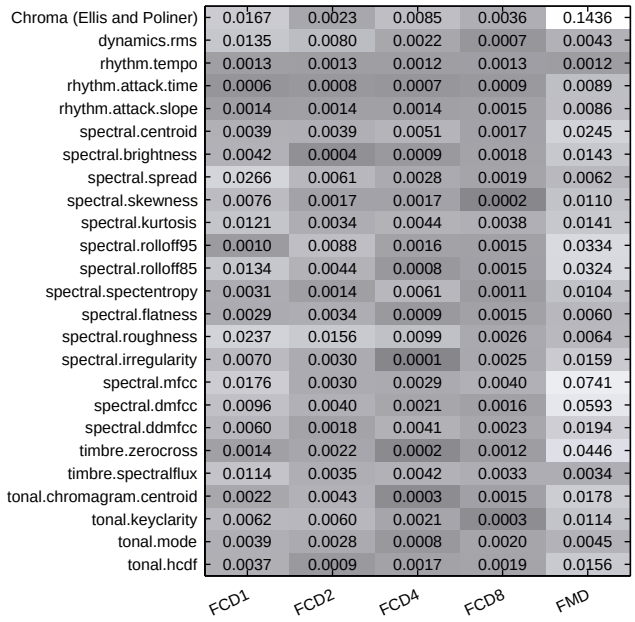


Figure 5.11: Normalised regression coefficient magnitudes, estimated using elastic net regularisation, for task of song year prediction. Candidate descriptor set comprised of FCDs and FMDs.

relatively strongly across multiple temporal scales. Note from Table 5.5 that for chroma features, MFCCs and derivatives, FMD weights are summed across 24 prediction coefficients, compared to 3 coefficients for FCDs.

In Figure 5.12 we examine prediction accuracy in response to windowed descriptors, as described in Equation 5.8 and where we quantify prediction accuracy 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. By contrast, for window sizes in the range $[0\text{d}\,..\,60\text{d}]$ using FMDs alone in place of FCDs alone we observe a mean performance gain of 7.0%.

5.4 Conclusion

In this chapter, 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-based information retrieval tasks, namely similarity rating prediction and song year prediction. We propose to compute track-wise sequential complexity as a descriptor for quantifying musical similarity. Whereas bag-of-features approaches such as feature moments disregard the temporal order of features, we conceive our descriptors as

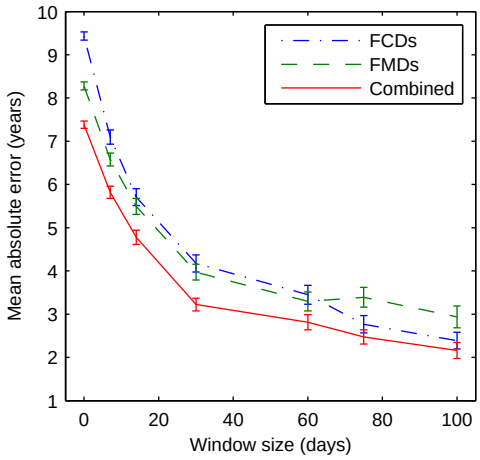


Figure 5.12: Song year prediction accuracy obtained using windowed descriptors, in response to window size. Error bars denote standard errors.

a statistic quantifying the amount of temporal regularity in a feature sequence.

Our proposed FCDs are computed in an unsupervised manner and may be implemented efficiently, requiring $O(n)$ time complexity for a feature sequence of length n (Effros, 2000). In addition, FCDs have similar dimensionality compared to FMDs. Since FCDs may be computed off-line or incrementally and then combined with efficient retrieval methods as considered by Slaney and Casey (2008); Rhodes et al. (2010); Schlüter (2013), we deem them potentially applicable in large-scale content-based information retrieval systems.

For both considered tasks, we observe that FCDs predict the outcome variable. Furthermore, in combination with FMDs, FCDs improve prediction accuracy with respect to the using FMDs alone. 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 predictive uncertainty may be used to improve the accuracy of low-specificity content-based information retrieval relying on bag-of-features approaches. Similar to results obtained by Foucard et al. (2011); Hamel et al. (2011, 2012); Dieleman and Schrauwen (2013), from examining estimated regression coefficients our results using FCDs suggest that an approach based on multiple temporal resolutions is advantageous for determining musical similarity.

Since FCDs by design relate to sequences of feature vectors rather than their marginal distribution, we expect that FCDs may complement FMDs (cf. Section 5.2). Whereas we indeed observe that FCDs combined with FMDs improve prediction accuracy compared to using FMDs alone, in addition we observe that FCDs by themselves do not outperform feature moments. Concerning the latter observation, we emphasise that as a measure of predictability, FCDs abstract

away from the values of feature vectors on which they are computed. That FMDs in isolation outperform FCDs in isolation by a maximum of 19.3% (cf. Section 5.3.1) thus appears unexpected. The performance of FCDs in isolation notwithstanding, with a view to applications in musical similarity we conclude that FCDs should be applied in combination with feature moments.

Chapter 6

Conclusions

6.1 Introduction

This chapter summarises and concludes this thesis. In Section 6.2 we summarise our investigations in chapters 4 and 5. We then relate the investigations to the research questions described in Chapter 1 and discuss ideas for future investigations. We subsequently prioritise ideas for future investigations in Section 6.4. Finally, we provide concluding remarks in Section 6.5.

6.2 Summary

In Chapter 4, we investigated using information-theoretic measures of predictability for cover song identification. Our evaluated measures quantify pairwise predictability between chroma feature sequences; we use such measures to determine pairwise similarity between tracks for the considered task. We evaluate discrete-valued and continuous-valued approaches as alternatives to the normalised compression distance (NCD).

Among discrete-valued methods, based on theoretical considerations, we propose normalised compression distance with alignment (NCDA) as an alternative to the NCD. Further, we propose methods based on prediction, which contrast with those based on string compression used to estimate NCD and NCDA. Secondly, we propose analogous methods directly applicable to continuous-valued feature sequences; thus no quantisation step is required to compute our measures when determining pairwise similarities between tracks. In the continuous-valued case, we compute our measures as statistics of the prediction error.

As we observe in our investigations on cover song identification, our proposed continuous-valued approaches consistently outperform discrete-valued approaches and compete with continuous-valued baseline approaches. Further, in the continuous-valued case, we observe that cross-prediction yields superior performance to our continuous-valued analogue of the normalised information distance. Concerning discrete-valued approaches, we observe that using NCDA instead of NCD improves cover song identification accuracy, for the case of the Lempel-Ziv (LZ) algorithm. Finally, we observe that continuous-valued prediction measures may be combined to improve performance relative to the baseline, attaining state-of-the-art performance using the Million Song Dataset (MSD).

In Chapter 5, we investigated using information-theoretic measures of predictability for similarity rating prediction and song year prediction. For this purpose, we quantify predictive uncertainty in track-wise manner. Our proposed feature complexity descriptors (FCDs) are based on computing predictive uncertainties at multiple time scales. Viewed as summary statistics in contrast to feature moments, our descriptors quantify temporal regularity, whereas feature moments disregard the temporal order of features.

For both similarity rating prediction and song year prediction, we observe that FCDs predict respective outcome variables. We observe that FCDs in isolation do not outperform the considered feature moment descriptors (FMDs). However, combined with FMDs, FCDs improve performance, compared to using FMDs alone. Further, our results suggest that using multiple time scales is advantageous for determining musical similarity.

Common to our investigations on cover song identification, similarity rating prediction and song year prediction, we observe that by including our proposed measures in predictive models, we improve accuracy relative to baselines, as quantified using our chosen performance statistics.

6.3 Discussion of Research Questions and Future Work

RQ3: To what extent can information-theoretic measures of predictability be used to determine musical similarity?

Recalling the research questions stated in Chapter 1, we begin by discussing research question **RQ3**, by which we subsequently infer answers about research questions **RQ1** and **RQ2**.

Examining our results on cover song identification, continuous-valued approaches attain competitive performance, compared to baseline approaches. This observation holds for both

the considered Jazz dataset and the MSD: quantified using mean average precision (MAP) and for the Jazz dataset, using our cross-prediction distance measure $D^{\times}$ we obtain an MAP score of 0.454, compared to MAP scores of 0.459 and 0.465 using cross-prediction normalised mean squared error (NMSE) and cross-correlation baselines, respectively. For the MSD, using our distance measure $D^{\times}$ we obtain an MAP score of 0.0498, compared to MAP scores of 0.0499 and 0.0404 for the same baseline approaches.

Further, we observe that we may combine our distance measures with baseline approaches to obtain significant performance gains with respect to baselines: for the Jazz dataset and MSD, by combining $D^{\times}$ with NMSE we obtain respective MAP scores of 0.496 and 0.0516, corresponding to performance gains of 8.1% and 3.4%. Considering the approaches proposed by Bertin-Mahieux and Ellis (2012), Khadkevich and Omologo (2013) involving the MSD, who respectively obtain MAP scores 0.0295, 0.0371, we obtain state-of-the-art results.

Concerning our evaluated low-specificity tasks, for similarity rating prediction we observe significant correlations between pairwise similarity ratings and pairwise distances between FCDs. These observations hold across the majority of evaluated features and considered time scales. Moreover, by incorporating FCDs into a multinomial regression model, quantified in terms of Spearman rank correlation and based on a four-point rating scale we obtain we obtain a relative performance gain of 31.1%, compared to using FMDs alone to predict similarity ratings. For song year prediction, by incorporating FCDs into a linear regression model we observe a relative performance gain of 10.9% compared to using FMDs alone to predict chart entry dates.

A limitation of our investigations on cover song identification is that we consider a single performance measure, the MAP, which we motivate to enable straightforward comparison between a multitude of approaches. While widely applied in version identification (cf. Serrà, 2011), it might prove instructive to consider alternative performance measures, such as mean reciprocal rank or precision at rank k (cf. Metzler, 2011). A further alternative might involve determining thresholds for identifying cover versions and computing precision/recall statistics, as performed by Casey et al. (2008a). Further, we combine distances using a single approach. For the latter problem of combining distances, further investigations might involve using artificial neural networks (ANNs) for regression (cf. Bishop, 2006).

For similarity rating prediction, 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. In both cases, a limitation of our work is that audio production may confound the association between similarity measures and respective outcome variables, as observed by Sturm (2012). While we submit the contending view that audio production is itself a musical characteristic, for future work it would be instructive to measure the influence of audio production by introducing suitable audio degradations, for example those proposed by Mauch and Ewert (2013). Extending investigations in Chapter 5 involving FCDs and FMDs in respective isolation, additional investigations might examine robustness of FCDs and FMDs in response to audio degradations. Further investigations might characterise the feature space relevant for similarity judgements.

RQ2: Is it possible to use information-theoretic measures of predictability to represent temporal structure in music?

From our obtained results, we conclude that our methods may be used to represent temporal structure in music: for cover song identification, all evaluated methods quantify similarity between sequences relevant for the task. In addition, for similarity rating prediction and song year-prediction our investigations confirm that temporal structure is relevant.

RQ1: How may we obtain representations of temporal structure in music?

As outlined in Chapter 1, for cover song identification we distinguish between discrete-valued and continuous-valued approaches, which we consider alternatives for obtaining representations of temporal structure in music. Comparing such methods, across considered datasets we observe that both our continuous cross-prediction measure $D^{\times}$ and the continuous baseline approaches (cross-prediction NMSE and cross-correlation) outperform all considered discrete-valued approaches. From these results we conclude that there are practical issues related to representing audio features using strings, and that the K-means based approach is insufficient. For future work, it would be beneficial to investigate alternative quantisation schemes, such as those based on chord transcription evaluated by Martin et al. (2012); Khadkevich and Omologo (2013). We view the continuous-valued approach as advantageous, since it requires no quantisation.

In both discrete-valued and continuous-valued cases, future investigations should examine the use of additional sequence modelling techniques. Among discrete-valued approaches, Shalizi and Shalizi (2004) propose a method for estimating hidden Markov models (HMMs) which

serve as optimal statistical predictors of sequences. The inferred models may subsequently be used to estimate entropy rates. An expansion of this approach has been proposed for the case of continuous-valued observations with clustering behaviour (Kelly et al., 2012). The latter approach incorporates a quantisation scheme which assigns a null symbol to ambiguous observations; null symbols are subsequently ignored during model estimation. We suggest that such an approach might be applied to chroma sequences. A further possibility for future investigations is the use of recurrent neural network architectures, such as long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), which we discussed in Chapter 2.

For the purpose of computing FCDs, our proposed use of a discrete-valued approach is pragmatic, based on the low computational cost of compressing sequences in the considered dataset. For future work, as proposed for cover song identification we might use HMMs to estimate entropy rates; furthermore as proposed continuous-valued prediction may readily be used to estimate entropy rates. Finally, it is conceivable that by using multiple approaches we might obtain more accurate estimates of entropy rates, using a suitable pooling function. Future investigations might examine whether more accurate estimates lead to improved similarity predictions.

RQ4: Which measures of predictability are useful for determining musical similarity?

From our investigations on cover song identification, we observe that for the LZ algorithm, NCD is outperformed by the proposed NCDA for both datasets, in terms of average performance gain across codebook sizes. Furthermore, NCD is outperformed by the proposed distance $D^{\times}$ based on discrete-valued cross-prediction and using the Jazz dataset. Thus, we conclude that among information-theoretic approaches, the NCD does not yield empirically optimal performance in general. It may therefore be advantageous to use alternative measures when faced with similar tasks.

In our present work on similarity rating prediction and song year prediction, we consider only a single measure of predictability. For further work, it would be instructive to determine the utility of alternative measures, for example those proposed by Dubnov (2008); Abdallah and Plumbley (2009). Since such measures are not functions of the entropy rate (cf. James et al., 2011), we suggest that they might be useful in combination with the proposed FCDs, potentially yielding further performance gains.

RQ5: Which feature representations are useful for determining musical similarity?

Our results using FCDs suggest that an approach based on multiple temporal resolutions is advantageous for determining musical similarity. This conclusion is based on comparing estimated regression coefficients across time scales used to represent feature sequences.

As an alternative to downsampled features, we initially employed beat-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 be influenced by short-term structure (such as timbral characteristics, as opposed to long-term structures such as motifs and chord progressions). For future work, we propose to examine in closer detail the utility of representing features at multiple time scales and to further characterise the feature space relevant for similarity judgements.

RQ6: How may we quantify similarity between sequences?

For the general problem of comparing strings, we propose NCDA as an alternative to the NCD. We motivate NCDA by considering sequences assumed to have been emitted by Markov sources; as proposed, NCDA is a modification of the NCD which accounts for correlation between sequences. Based on artificial sequences, we demonstrate experimentally that NCDA more closely than NCD approximates the normalised information distance (NID), the latter which we estimate analytically. Thus, we conclude that NCDA may yield more favourable performance in tasks where we may assume Markov sources underlying observed sequences. Our results on cover song identification support this conclusion based on consistent improvements using NCDA versus NCD, for the LZ algorithm. However, compared to experiments involving artificial strings, we observe inconsistent results for Burrows-Wheeler (BW) and prediction by partial matching (PPM) algorithms.

With the aim of establishing causes for differences between results involving artificial strings and quantised chroma sequences, future investigations should examine in detail the behaviour of NCDA when applied to BW and PPM algorithms. To this end, further analysis should quantify the performance improvement of NCDA applied to individual compression steps in each respec-

tive algorithm. The artificial strings with equi-probable binary observations and noise based on an independent, identically distributed random variable (as considered in Chapter 4) might subsequently be extended to more ecologically valid data such as annotated chord sequences, the latter which might be combined with a more realistic noise source based on probabilistic insertion, substitution and deletion. Further, future investigations should examine in greater detail the behaviour of the string alignment step in NCDA, both as proposed in Chapter 4 and substituted with alternatives such as dynamic time warping (DTW).

Future work might also examine the behaviour of NCDA versus NCD applied to alternative tasks to cover song identification. Similarly, future investigations might involve applying our proposed continuous-valued approaches to alternative tasks.

RQ7: How might we perform computationally efficient retrieval?

Our evaluations of pairwise and track-wise measures of predictability involve metric spaces for computing similarity: for similarity rating prediction, we compute Euclidean distances between FCDs, whereas for song year prediction we apply linear regression to FCD values themselves. For cover song identification, we incorporate our methods of computing pairwise predictability into a two-stage process, where in the first stage we filter tracks using the L1 distance. By expressing similarity using a metric, in all cases we allow for the possibility of applying techniques with sub-linear retrieval time complexity, with respect to the number of tracks in the collection (cf. Slaney and Casey, 2008).

While we allow for the possibility of using such approaches, it remains to examine retrieval performance using commercial-scale datasets. Future work should first examine retrieval accuracy when using sub-linear techniques, as considered by Schlüter (2013): given a specified collection size it may be nevertheless advantageous to perform a linear scan, providing that pairwise comparisons can be performed with sufficient computational efficiency. Future work should evaluate retrieval performance based on performing linear scans, using optimised implementations of our methods.

6.4 Priority List of Future Investigations

Based on ideas for future work discussed in the preceding section, we prioritise future investigations of immediate importance as follows:

1. Establish causes for differences in performance gains, observed when computing NCDA across compressors and across considered string data.

With a view to establishing causes for differences in performance gains when computing NCDA across LZ, BW and PPM compressors, as a starting point we propose to restrict investigations to simplified implementations of respective compressors. For LZ compression, investigations should involve the algorithm originally proposed by Ziv and Lempel (1977), whereas for PPM investigations should involve the algorithm proposed by Cleary and Witten (1984). For BW compression, investigations should involve the Burrows-Wheeler transform (Burrows and Wheeler, 1994).

The effect of string interleaving should be investigated using theoretical analysis, artificial data, as well as real-world data. After illustrating algorithm behaviour using example strings, theoretical analysis should serve as a starting point for distinguishing among the theoretical behaviour of NCDA for the considered compression algorithms. Thereafter, experiments involving artificial data should be used to test hypothesised behaviour obtained from theoretical analysis, as well as expand on the empirical behaviour of NCDA. As sources for artificially generated data, we propose the use of Markov and golden mean processes (cf. James et al., 2011). As a noise source, we propose the use of a probabilistic substitution table, with substitution probabilities sampled from a Dirichlet process. Both artificial and real-word data should subsequently be used to evaluate the behaviour of complex compression techniques, with the aim of more closely matching the behaviour of general-purpose compressors. In complex compression techniques, the effect of string interleaving should be evaluated with respect to combinations of constituent compression steps. The proposed investigations contrast with those described in Chapter 4, the latter which evaluate NCDA in combination with general-purpose compressors.

2. Evaluate NCDA using alternative alignment techniques.

Having evaluated the effect of string interleaving across compressors, future investigations should establish how the performance of NCDA is influenced by the chosen string alignment technique. To this end, we propose to compute distances between string pairs, respectively transformed with varying amounts of probabilistically occurring symbol insertions, substitutions and deletions. We would then aim at contrasting our proposed method against alternatives based on DTW, seeking to quantify the amount of robustness in NCDA with respect to string transformations. A starting point for modifying our proposed method might involve performing correlation-based

alignment at the level of unquantised features; it would subsequently be of interest to compare the computational complexity of competing approaches. Another variant of correlation-based alignment applicable to non-binary alphabets might incorporate a binary comparison function between symbols. Finally, since DTW may be viewed as an alternative distance measure to NCD, it would be instructive to compare the performance of the two respective approaches and to explore the possibility of combining distances obtained using DTW and NCDA.

3. Evaluate alternative quantisation techniques for obtaining discrete-valued representations.

Further investigations should establish how the performance of NCDA and NCD is influenced by the chosen quantisation method. To this aim, we propose to contrast K-means against the chord transcription methods proposed by Mauch (2010) and subsequently implemented as the software Chordino [1]. Further contrasting with K-means, we propose to evaluate alternative encoding strategies based on computing cosine similarities with respect to codewords (cf. Vaizman et al., 2014); extended to the problem of codebook learning we note the possibility of using spherical K-means clustering, where codewords are normalised to unit L2 norm (cf. Coates and Ng, 2012). For the case of chroma features, we view an approach based on cosine similarities as potentially advantageous compared to Euclidean K-means, since it allows us to incorporate chroma feature normalisation in the codebook learning or codeword assignment steps. Investigations should establish whether performance gains can be achieved, compared to our approach of normalising feature vectors prior to codebook learning and quantisation.

4. Investigate methods for combining distances.

The method for combining pairwise distances considered in Chapter 4 represents a starting point towards a more detailed investigation on using ensemble techniques for cover song identification. Retaining the approach of using ranked distances as the input feature space, we propose to investigate training ANNs, using MAP as the loss function which we seek to minimise. A potential difficulty using MAP is that training ANNs may require a continuous loss function with respect to network parameters. Investigations should alternatively consider cover song identification as an ordinal regression problem, as well as a classification problem. Thus formulated, a further supervised learning approach might involve support vector machines (SVMs). In the work of Ravuri and Ellis (2010), we note an approach based on combining distances using SVMs and ANNs for classification.

[1] `http://isophonics.net/nnls-chroma`, retrieved April 2015.

5. Examine behaviour of FCDs in response to audio transformations

Further investigations should establish in detail the feature spaces represented by FCDs. To this aim, as previously described we may degrade musical audio using suitable transformations (cf. Mauch and Ewert, 2013). We propose an exploratory analysis, whereby we contrast FCDs against FMDs, with respect to varying amounts of dynamic range compression, varying amounts of reverberation and varying tempo. We note that depending on the intended application, we may consider tempo variation both an irrelevant degradation, as well as a musically relevant transformation. Thus, we propose to quantify the behaviour of FCDs versus FMDs both in terms of robustness, as well as variability of descriptors with respect to transformations. With a view to examining in further detail the use of multiple time scale representations, the proposed investigations should involve FCDs computed using downsampled feature sequences, as proposed in Chapter 5.

6.5 Conclusion

This thesis has investigated the use of information-theoretic measures of predictability for music content analysis tasks. In particular, we have examined using information-theoretic methods as a conceptual framework for obtaining representations of temporal structure in music, for the purpose of determining musical similarity. We have demonstrated that our approach benefits music content analysis tasks based on musical similarity.

Recalling the properties of abstraction, generality and expressiveness identified in Chapter 2, we offer the following conclusions: abstraction has allowed us to compare methods applicable to diverse feature representations, namely discrete-valued and continuous-valued representations, obtained at diverse time scales. Generality has allowed us to consider a variety of methods for estimating our measures. Finally, expressiveness has allowed us to consider a variety of possible measures. Taken together, these properties have enabled an inquiry into how we might determine musical similarity. Our obtained results and conclusions convey utility in our framework.

There exists substantial scope for expanding on the sequential models evaluated in this thesis. A limitation of the present work is that single tracks or pairs of tracks are modelled in isolation. Future work might examine the potential of quantifying predictability using a global model, akin to the ‘long-term model’ in the work of Conklin and Witten (1995); Pearce and Wiggins (2006) involving symbolic representations. Such an approach might be useful for modelling subjective

notions of similarity, which we did not investigate in this work.

Furthermore, the measures which we compute are summary statistics with respect to entire tracks, or pairs of tracks. Thus, we model sequences in their entirety, rather than incrementally. Future work might investigate the alternative of an information-dynamic approach, where predictive uncertainty is quantified with respect to incremental models (cf. Abdallah and Plumbley, 2009). It would then be of interest to determine whether the resulting information profiles may be used to determine musical similarity.

Finally, the possibility of using ensemble techniques demands attention. Besides combining at the level of measures as previously suggested, it may prove useful to combine at the level of individual predictions. To this end, we might consider online ensemble methods (cf. Vovk, 2001; Cesa-Bianchi and Lugosi, 2006). It would subsequently be of interest to determine whether improved estimates of information-theoretic quantities relate to improved performance at similarity tasks. Ensemble techniques might further be applied across multiple feature representations: for the purpose of version identification, it may prove fruitful to predict chroma features using continuous-valued approaches in addition to discrete-valued approaches, with subsequent conversion to a common representation.

Bibliography

B. Aarden. *Dynamic Melodic Expectancy*. PhD thesis, Ohio State University, United States of America, 2003.

S. A. Abdallah and M. D. Plumbley. Information dynamics: Patterns of expectation and surprise in the perception of music. *Connection Science*, 21(2):89–117, 2009.

S. A. Abdallah and M. D. Plumbley. Predictive information in Gaussian processes with application to music analysis. In *Geometric Science of Information*, pages 650–657. Springer, 2013.

A. Agresti. *Analysis of Ordinal Categorical Data*. John Wiley and Sons, 2010.

T. E. Ahonen. Measuring harmonic similarity using PPM-based compression distance. In *Proceedings of the 1st Workshop Exploring Musical Information Spaces (WEMIS)*, pages 52–55, 2009.

T. E. Ahonen. Combining chroma features for cover version identification. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 165–170, 2010.

T. E. Ahonen. Compression-based clustering of chromagram data: New method and representations. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, pages 474–481, 2012.

T. E. Ahonen, K. Lemström, and S. Linkola. Compression-based similarity measures in symbolic, polyphonic music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 91–96, 2011.

C. Allauzen, M. Crochemore, and M. Raffinot. Factor oracle: A new structure for pattern matching. In *SOFSEM'99: Theory and Practice of Informatics*, pages 295–310, 1999.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

A. Anglade, E. Benetos, M. Mauch, and S. Dixon. Improving music genre classification using automatically induced harmony rules. *Journal of New Music Research*, 39(4):349–361, 2010.

J. A. Aslam, E. Yilmaz, and V. Pavlu. A geometric interpretation of R-precision and its correlation with average precision. In *Proceedings of the 28th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 573–574, 2005.

G. Assayag, S. Dubnov, and O. Delerue. Guessing the composer's mind: Applying universal prediction to musical style. In *Proceedings of the 25th International Computer Music Conference (ICMC)*, pages 496–499, 1999.

J.-J. Aucouturier and F. Pachet. Music similarity measures: What's the use? In *Proceedings of the 3rd International Society for Music Information Retrieval Conference (ISMIR)*, pages 157–163, 2002.

J.-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. *Journal of the Acoustical Society of America*, 122(2):881–891, 2007.

G. J. Balzano and B. W. Liesch. The role of chroma and scalestep in the recognition of musical intervals in and out of context. *Psychomusicology: Music, Mind and Brain*, 2(2):3–31, 1982.

M. A. Bartsch and G. H. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *Proceedings of the 4th IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 15–18, 2001.

R. Begleiter, R. El-Yaniv, and G. Yona. On prediction using variable order Markov models. *Journal of Artificial Intelligence Research*, 22:385–421, 2004.

J. P. Bello. Audio-based cover song retrieval using approximate chord sequences: Testing shifts, gaps, swaps and beats. In *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*, pages 239–244, 2007.

J. P. Bello. Measuring structural similarity in music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2013–2025, 2011.

Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu. Advances in optimizing recurrent networks. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8624–8628, 2013.

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*, 28(2):63–76, 2004.

T. Bertin-Mahieux and D. P. W. Ellis. Large-scale cover song recognition using hashed chroma landmarks. In *Proceedings of the 9th IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 117–120, 2011.

T. Bertin-Mahieux and D. P. W. Ellis. Large-scale cover song recognition using the 2D Fourier transform magnitude. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 241–246, 2012.

T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 591–596, 2011.

J. J. Bharucha and P. M. Todd. Modeling the perception of tonal structure with neural nets. *Computer Music Journal*, 13(4):44–53, 1989.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 2007.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

D. Bogdanov, J. Serrà, N. Wack, P. Herrera, and X. Serra. Unifying low-level and high-level music similarity measures. *IEEE Transactions on Multimedia*, 13(4):687–701, 2011.

T. L. Bolton. Rhythm. *The American Journal of Psychology*, 6(2):145–238, 1894.

G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. John Wiley and Sons, 2013.

F. P. Brooks, A. L. Hopkins, P. G. Neumann, and W. V. Wright. An experiment in musical composition. *IRE Transactions on Electronic Computers*, EC-6(3):175–182, 1957.

A. E. Bryson, W. F. Denham, and S. E. Dreyfus. Optimal programming problems with inequality constraints. *AIAA Journal*, 1(11):2544–2550, 1963.

M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. Technical report, Digital Equipment Corporation, 1994.

D. Byrd. A similarity scale for content-based music IR. `http://www.informatics.indiana.edu/donbyrd/MusicSimilarityScale.HTML`, 2007. Accessed: 19.08.14.

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*, 41(3):271–284, 2005.

J. S. Cardoso and R. Sousa. Measuring the performance of ordinal classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(8):1173–1195, 2011.

J. C. Carlsen, P. I. Divenyi, and J. A. Taylor. A preliminary study of perceptual expectancy in melodic configurations. *Bulletin of the Council for Research in Music Education*, pages 4–12, 1970.

M. A. Casey and M. Slaney. The importance of sequences in musical similarity. In *Proceedings of the 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 5–8, 2006.

M. A. Casey, C. Rhodes, and M. Slaney. Analysis of minimum distances in high-dimensional musical spaces. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):1015–1028, 2008a.

M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96 (4):668–696, 2008b.

M. Cebrian, M. Alfonseca, and A. Ortega. Common pitfalls using the normalized compression distance: What to watch out for in a compressor. *Communications in Information and Systems*, 5(4):367–384, 2005.

O. Celma. *Music Recommendation and Discovery in the Long Tail*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2009.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

S. Cherla, T. Weyde, A. S. d'Avila Garcez, and M. T. Pearce. A distributed model for multiple-viewpoint melodic prediction. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 15–20, 2013.

R. Cilibrasi and P. M. B. Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005.

M. Clausen and F. Kurth. A unified approach to content-based and fault-tolerant music recognition. *IEEE Transactions on Multimedia*, 6(5):717–731, 2004.

J. G. Cleary and I. Witten. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, 32(4):396–402, 1984.

A. Coates and A. Y. Ng. Learning feature representations with K-means. In *Neural Networks: Tricks of the Trade*, pages 561–580. Springer, 2012.

J. E. Cohen. Information theory and music. *Behavioral Science*, 7(2):137–163, 1962.

D. Conklin and J. G. Cleary. Modelling and generating music using multiple viewpoints. In *Proceedings of the 1st Workshop on Artificial Intelligence and Music*, pages 125–137, 1988.

D. Conklin and I. H. Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73, 1995.

A. Cont. *Modeling Musical Anticipation: From the Time of Music to the Music of Time*. PhD thesis, University of California, San Diego, United States of America; Université Pierre et Marie Curie, Paris, France, 2008.

E. Coons and D. Kraehenbuehl. Information as a measure of structure in music. *Journal of Music Theory*, pages 127–161, 1958.

T. M. Cover and J. A.. Thomas. *Elements of Information Theory*. John Wiley and Sons, 2012.

E. Coviello, Y. Vaizman, A. B. Chan, and G. R. G. Lanckriet. Multivariate autoregressive mixture models for music auto-tagging. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 547–552, 2012.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

I. Cross. Music, cognition, culture, and evolution. *Annals of the New York Academy of Sciences*, 930(1):28–42, 2001.

M. E. Curtis and J. J. Bharucha. Memory and musical expectation for tones in cultural context. *Music Perception*, 26(4):365–375, 2009.

A. de Cheveigne. *Pitch*. Springer, 2005.

P. Desain, H. Honing, and M. Sadakata. Predicting rhythm perception from rhythm production and score counts: The Bayesian approach. In *Proceedings of the 6th Society for Music Perception and Cognition Conference (SMPC)*, 2003.

S. Dieleman and B. Schrauwen. Multiscale approaches to music audio feature learning. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 3–8, 2013.

S. Dieleman, P. Brakel, and B. Schrauwen. Audio-based music classification with a pretrained convolutional network. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 669–674, 2011.

S. Dixon. A lightweight multi-agent musical beat tracking system. In *Pacific Rim International Conference on Artificial Intelligence (PRICAI): Topics in Artificial Intelligence*, pages 778–788. Springer, 2000.

S. Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1):39–58, 2001.

J. S. Downie. Music information retrieval. *Annual Review of Information Science and Technology*, 37(1):295–340, 2003.

J. S. Downie, M. Bay, A. F. Ehmann, and M. C. Jones. Audio cover song identification: MIREX 2006–2007 results and analyses. In *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR)*, pages 468–474, 2008.

J. S. 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*, pages 93–115. Springer, 2010.

S. Dubnov. Spectral anticipations. *Computer Music Journal*, 30(2):63–83, 2006.

S. Dubnov. Unified view of prediction and repetition structure in audio signals with application to interest point detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):327–337, 2008.

S. Dubnov, S. McAdams, and R. Reynolds. Structural and affective aspects of music from statistical audio signal analysis. *Journal of the American Society for Information Science and Technology*, 57(11):1526–1536, 2006.

S. Dubnov, G. Assayag, and A. Cont. Audio oracle: A new algorithm for fast learning of audio structures. In *Proceedings of the 33rd International Computer Music Conference (ICMC)*, 2007.

D. Eck and J. Schmidhuber. Finding temporal structure in music: Blues improvisation with LSTM recurrent networks. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, pages 747–756, 2002.

T. Eerola. Data-driven influences on melodic expectancy: Continuations in North Sami yoiks rated by South African traditional healers. In *Proceedings of the 8th International Conference of Music Perception and Cognition (ICMPC)*, pages 83–87, 2004.

M. Effros. PPM performance with BWT complexity: A new method for lossless data compression. In *Proceedings of the 10th Data Compression Conference (DCC)*, pages 203–212, 2000.

B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, 1982.

D. P. W. Ellis. Beat tracking with dynamic programming. *Proceedings of the Music Information Retrieval Evaluation Exchange (MIREX)*, 2006.

D. P. W. Ellis. Classifying music audio with timbral and chroma features. In *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*, pages 339–340, 2007.

D. P. W. Ellis and G.E. Poliner. Identifying 'cover songs' with chroma features and dynamic

programming beat tracking. In *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1429–1432, 2007.

C. Faloutsos and K. I. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the 26th ACM International Conference on Management of Data (SIGMOD)*, pages 163–174, 1995.

M. Feder and N. Merhav. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40(1):259–266, 1994.

M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38(4):1258–1270, 1992.

D. P. Feldman. A brief introduction to: Information theory, excess entropy and computational mechanics. Technical report, College of the Atlantic, United States of America, 2002.

P. Fenwick. Block sorting text compression. *Australian Computer Science Communications*, 18: 193–202, 1996.

A. Flexer and D. Schnitzer. Effects of album and artist filters in audio similarity computed for very large music databases. *Computer Music Journal*, 34(3):20–28, 2010.

J. T. Foote. Content-based retrieval of music and audio. In *Proceedings of SPIE Multimedia Storage and Archiving Systems II*, pages 138–147, 1997.

J. T. Foote. ARTHUR: Retrieving orchestral music by long-term structure. In *Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR)*, 2000.

P. Foster, A. Klapuri, and M. D. Plumbley. Causal prediction of continuous-valued music features. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 501–506, 2011.

P. Foster, S. Dixon, and A. Klapuri. Identification of cover songs using information theoretic measures of similarity. In *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 739–743, 2013.

P. Foster, A. Klapuri, and S. Dixon. Identifying cover songs using information-theoretic measures of similarity. *arXiv preprint arXiv:1407.2433*, 2014a.

P. Foster, M. Mauch, and S. Dixon. Sequential complexity as a descriptor for musical similarity. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1967–1977, 2014b.

R. Foucard, S. Essid, M. Lagrange, and G. Richard. Multi-scale temporal fusion by boosting for music classification. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 663–668, 2011.

P. Fraisse. Rhythm and tempo. In *The Psychology of Music*, pages 149–180. Academic Press, 1982.

J. P. Friedlander. News and notes on 2014 mid-year RIAA shipment and revenue statistics. Technical report, Recording Industry Association of America, 2014.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

Z. Fu, G. Lu, K. M. Ting, and D. Zhang. Music classification via the bag-of-features approach. *Pattern Recognition Letters*, 32(14):1768–1777, 2011a.

Z. Fu, G. Lu, K. M. Ting, and D. Zhang. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2):303–319, 2011b.

T. Fujishima. Realtime chord recognition of musical sound: A system using Common Lisp Music. In *Proceedings of the 25th International Computer Music Conference (ICMC)*, pages 464–467, 1999.

R. G. Gallager. *Principles of Digital Communication*. Cambridge University Press, 2008.

E. Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2006.

E. Gómez and P. Herrera. The song remains the same: Identifying versions of the same piece using tonal descriptors. In *Proceedings of the 7th International Society for Music Information Retrieval Conference (ISMIR)*, 2006.

N. Goodman. Seven strictures on similarity. In *Problems and Projects*. Bobbs-Merrill, 1972.

F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer. Evaluating rhythmic descriptors for musical genre classification. In *Proceedings of the 25th International AES Conference*, pages 196–204, 2004.

C. W. J. Granger and R. Joyeux. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1(1):15–29, 1980.

G. Z. Greenberg and W. D. Larkin. Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe-signal method. *Journal of the Acoustical Society of America*, 44(6):1513–1523, 1968.

P. Grünwald and P. M. B. Vitányi. Shannon information and Kolmogorov complexity. *arXiv preprint arXiv:cs/0410002*, 2004.

P. Hamel, S. Lemieux, Y. Bengio, and D. Eck. Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 729–734, 2011.

P. Hamel, Y. Bengio, and D. Eck. Building musically-relevant audio features through multiple timescale representations. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 553–558, 2012.

E. Hanslick. *On the Musically Beautiful: A Contribution Towards the Revision of the Aesthetics of Music (translated by G. Payzant)*. Hackett Publishing (Original work published 1891), 1986.

C. Harte, M. Sandler, and M. Gasser. Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, pages 21–26, 2006.

A. Hazan. *Musical Expectation Modelling from Audio: A Causal Mid-level Approach to Predictive Representation and Learning of Spectro-temporal Events*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2010.

M. Helén and T. Virtanen. Audio query by example using similarity measures between probability density functions of features. *EURASIP Journal on Audio, Speech, and Music Processing*, DOI 10.1155/2010/179303, 2010.

W. E. Hick. On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4(1):11–26, 1952.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

G. E. Hinton. Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11(10): 428–434, 2007.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

J. H. Howard, A. J. O'Toole, R. Parasuraman, and K. B. Bennett. Pattern-directed attention in uncertain-frequency detection. *Perception and Psychophysics*, 35(3):256–264, 1984.

D. B. Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, 2006.

R. Hyman. Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, 45(3):188–196, 1953.

R. Jackendoff. *Consciousness and the Computational Mind*. MIT Press, 1987.

R. G. James, C. J. Ellison, and J. P. Crutchfield. Anatomy of a bit: Information in a time series observation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(3):037109, 2011.

T. Jehan. Analyzer documentation. Technical report, The Echo Nest, 2011.

J. H. Jensen, M. G. Christensen, and S. H. Jensen. A chroma-based tempo-insensitive distance measure for cover song identification using the 2D autocorrelation function. *Proceedings of the Music Information Retrieval Evaluation Exchange (MIREX)*, 2008.

K. Jensen. *Timbre Models of Musical Sounds*. PhD thesis, University of Copenhagen, Denmark, 1999.

M. R. Jones, H. Moynihan, N. MacKenzie, and J. Puente. Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science*, 13(4):313–319, 2002.

A. Kaltchenko. Algorithms for estimating information distance with application to bioinformatics and linguistics. In *Proceedings of the 17th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, volume 4, pages 2255–2258, 2004.

D. Kelly, M. Dillingham, A. Hudson, and K. Wiesner. A new method for inferring hidden Markov models from noisy time sequences. *PloS One*, 7(1):e29703, 2012.

M. Kennedy, J. Kennedy, and T. Rutherford-Johnson. *The Oxford Dictionary of Music*. Oxford University Press, 2013.

M. Khadkevich and M. Omologo. Large-scale cover song identification using chord profiles. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 233–238, 2013.

Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. Music emotion recognition: A state of the art review. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 255–266, 2010.

G. Kirchgassner, J. Wolters, and U. Hassler. *Introduction to Modern Time Series Analysis*. Springer, 2012.

A. Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proceedings of the 7th International Society for Music Information Retrieval Conference (ISMIR)*, pages 216–221, 2006.

C. L. Krumhansl. Rhythm and pitch in music cognition. *Psychological Bulletin*, 126(1):159, 2000.

F. Kurth and M. Müller. Efficient index-based audio matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, 2008.

T. Langlois and G. Marques. A music classification method based on timbral features. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 81–86, 2009.

E. W. Large. On synchronizing movements to music. *Human Movement Science*, 19(4):527–566, 2000.

O. Lartillot and P. Toiviainen. A Matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFX)*, pages 237–244, 2007.

C. Lee, J. Shih, K. Yu, and H. Lin. Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. *IEEE Transactions on Multimedia*, 11(4): 670–682, 2009.

K. Lee. Identifying cover songs from audio using harmonic representation. *Proceedings of the Music Information Retrieval Evaluation Exchange (MIREX)*, 2006.

M. Levy and M. Sandler. Lightweight measures for timbral similarity of musical audio. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, pages 27–36, 2006.

M. Li and R. Sleep. Melody classification using a similarity metric based on Kolmogorov complexity. In *Proceedings of the 1st Sound and Music Computing Conference (SMC)*, pages 126–129, 2004.

M. Li and R. Sleep. Genre classification via an LZ78-based string kernel. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, pages 252–259, 2005.

M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer, 2008.

M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.

T. Li and M. Ogihara. Toward intelligent music information retrieval. *IEEE Transactions on Multimedia*, 8(3):564–574, 2006.

B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR)*, 2000.

B. Logan and A. Salomon. A music similarity function based on signal analysis. In *Proceedings of the 2nd IEEE International Conference on Multimedia and Expo. (ICME)*, 2001.

J. London. *Hearing in Time*. Oxford University Press, 2012.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2005.

E. Mach. Untersuchungen über den Zeitsinn des Ohres. *Sitzungsberichte der Mathematisch-naturwissenschaftlichen Classe der Kaiserlichen Akademie der Wissenschaften*, 51(2):133–150, 1865.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

M. Mandel and D. P. W. Ellis. Song-level features and support vector machines for music classification. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, pages 594–599, 2005.

M. Marchini and H. Purwins. Unsupervised generation of percussion sound sequences from a sound example. In *Proceedings of the 7th Sound and Music Computing Conference (SMC)*, 2010.

M. Marolt. A mid-level representation for melody-based retrieval in audio collections. *IEEE Transactions on Multimedia*, 10(8):1617–1625, 2008.

B. Martin, D. G. Brown, P. Hanna, and P. Ferraro. BLAST for audio sequences alignment: A fast scalable cover identification tool. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 529–534, 2012.

R. Marxer and H. Purwins. Unsupervised incremental learning and prediction of audio signals. In *Proceedings of 20th International Symposium on Music Acoustics (ISMA)*, 2010.

P. Masri. *Computer Modelling of Sound for Transformation and Synthesis of Musical Signals*. PhD thesis, University of Bristol, United Kingdom, 1996.

M. Mauch. *Automatic Chord Transcription from Audio Using Computational Models of Musical Context*. PhD thesis, Queen Mary University of London, United Kingdom, 2010.

M. Mauch and S. Ewert. The audio degradation toolbox and its application to robustness evaluation. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 83–88, 2013.

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*, 58(3):177–192, 1995.

A. Meng and J. Shawe-Taylor. An investigation of feature models for music genre classification using the support vector classifier. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, pages 604–609, 2005.

A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen. Temporal feature integration for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1654–1664, 2007.

D. Metzler. *A Feature-centric View of Information Retrieval*. Springer, 2011.

L. B. Meyer. *Emotion and Meaning in Music*. University of Chicago Press, 1956.

F. Mörchen, A. Ultsch, M. Thies, and I. Lohken. Modeling timbre distance with temporal statistics from polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):81–90, 2006.

M. C. Mozer. Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multi-scale processing. *Connection Science*, 6(2-3):247–280, 1994.

S. Muggleton. Inductive logic programming. *New Generation Computing*, 8(4):295–318, 1991.

M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard. Signal processing for music analysis. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1088–1110, 2011.

E. Narmour. *The Analysis and Cognition of Basic Melodic Structures: The Implication-realization Model*. University of Chicago Press, 1990.

J. E. Ollen. *A Criterion-related Validity Test of Selected Indicators of Musical Sophistication Using Expert Ratings*. PhD thesis, Ohio State University, United States of America, 2006.

H. F. Olson. *Music, Physics and Engineering*. Courier Corporation, 1967.

E. Pampalk. *Computational Models of Music Similarity and Their Application in Music Information Retrieval*. PhD thesis, Vienna University of Technology, Austria, 2006.

A. D. Patel. *Music, Language, and the Brain*. Oxford University Press, 2010.

Jouni Paulus, M. Müller, and A. Klapuri. State of the art report: Audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 625–636, 2010.

M. T. Pearce. Early applications of information theory to music. Technical report, Goldsmiths, University of London, United Kingdom, 2007.

M. T. Pearce and G. A. Wiggins. Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, 33(4):367–385, 2004.

M. T. Pearce and G. A. Wiggins. Expectation in melody: The influence of context and learning. *Music Perception*, 23(5), 2006.

M. T. Pearce, M. H. Ruiz, S. Kapasi, G. A. Wiggins, and J. Bhattacharya. Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, 50(1):302–313, 2010.

R. C. Pinkerton. Information theory and melody. *Scientific American*, 194(2), 1956.

J. F. Pinto da Costa, H. Alonso, and J. S. Cardoso. The unimodal model for the classification of ordinal data. *Neural Networks*, 21(1):78–91, 2008.

W. Piston. *Harmony*. Norton, 1978.

R. Plomp and W. J. M. Levelt. Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America*, 38(4):548–560, 1965.

D. Ponsford, G. A. Wiggins, and C. Mellish. Statistical learning of harmonic movement. *Journal of New Music Research*, 28(2):150–177, 1999.

J. Pressing. Cognitive complexity and the structure of musical patterns. In *Proceedings of the 4th Australasian Cognitive Science Society Conference*, 1999.

J. Qian, T. Hastie, J. Friedman, R. Tibshirani, and N. Simon. Glmnet for Matlab. `http://www.stanford.edu/~hastie/glmnet_matlab/`, 2013. Accessed: 19.08.14.

L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. Searching and mining trillions of time series subsequences under dynamic time

warping. In *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 262–270, 2012.

R. Rasch and R. Plomp. The perception of musical tones. In *The Psychology of Music*, pages 89–112. Academic Press, 1999.

S. Ravuri and D. P. W. Ellis. Cover song detection: From high scores to general classification. In *Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 65–68, 2010.

J. Reed and C. Lee. On the importance of modeling temporal information in music tag annotation. In *Proceedings of the 34th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1873–1876, 2009.

J. Ren and J. Jang. Discovering time-constrained sequential patterns for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1134–1144, 2012.

B. H. Repp. Rate limits in sensorimotor synchronization with auditory and visual sequences: The synchronization threshold and the benefits and costs of interval subdivision. *Journal of motor behavior*, 35(4):355–370, 2003.

C. Rhodes, T. Crawford, M. A. Casey, and M. d'Inverno. Investigating music collections at different scales with AudioDB. *Journal of New Music Research*, 39(4):337–348, 2010.

M. A. Rohrmeier and S. Koelsch. Predictive information processing in music cognition. A critical review. *International Journal of Psychophysiology*, 83(2):164–175, 2012.

J. R. Saffran, E. K. Johnson, R. N. Aslin, and E. L. Newport. Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52, 1999.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

E. G. Schellenberg. Expectancy in melody: Tests of the implication-realization model. *Cognition*, 58(1):75–125, 1996.

J. Schlüter. Learning binary codes for efficient large-scale music similarity search. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 581–586, 2013.

J. Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Anticipatory Behavior in Adaptive Learning Systems*, pages 48–76. Springer, 2009.

D. Schnitzer, A. Flexer, and G. Widmer. A filter-and-refine indexing method for fast similarity search in millions of music tracks. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 537–542, 2009.

D. Sculley and C. E. Brodley. Compression and machine learning: A new perspective on feature space vectors. In *Proceedings of the 16th Data Compression Conference (DCC)*, pages 332–341, 2006.

J. Serrà. *Identification of Versions of the Same Musical Composition by Processing Audio Descriptions*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2011.

J. Serrà, E. Gómez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1138–1151, 2008.

J. Serrà, X. Serra, and R. G. Andrzejak. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9):093017, 2009.

J. Serrà, H. Kantz, X. Serra, and R. G. Andrzejak. Predictability of music descriptor time series and its application to cover song detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):514–525, 2012.

X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gómez, F. Gouyon, P. Herrera, S. Jordà, O. Paytuvi, G. Peeters, J. Schlüter, H. Vinet, and G. Widmer. Roadmap for music information research. Technical report, MIReS Consortium, 2013.

K. Seyerlehner, G. Widmer, and T. Pohle. Fusing block-level features for music similarity estimation. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFX)*, pages 225–232, 2010.

C. R. Shalizi and K. L. Shalizi. Blind construction of optimal nonlinear recursive predictors for discrete sequences. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 504–511, 2004.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, 1948.

C. E. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1): 50–64, 1951.

D. Silva, H. Papadopoulos, G.E.A.P.A Batista, and D. P. W. Ellis. A video compression-based approach to measure music structural similarity. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 95–100, 2013.

P. Skibinski and Si. Grabowski. Variable-length contexts for PPM. In *Proceedings of the 14th Data Compression Conference (DCC)*, pages 409–418, 2004.

M. Slaney. Auditory Toolbox Version 2. Technical report, Interval Research Corporation, 1998.

M. Slaney and M. A. Casey. Locality-sensitive hashing for finding nearest neighbors. *IEEE Signal Processing Magazine*, 25(2):128–131, 2008.

M. Slaney, K. Q. Weinberger, and W. White. Learning a metric for music similarity. In *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR)*, pages 313–318, 2008.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

A. M. Stark and M. D. Plumbley. Performance following: Real-time prediction of musical sequences without a score. *IEEE Transactions on Audio, Speech, and Language Processing*, 20 (1):190–199, 2012.

S. Streich. *Automatic Characterization of Music Complexity: A Multifaceted Approach*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2006.

B. L. Sturm. Two systems for automatic music genre recognition: What are they really recognizing? In *Proceedings of the 2nd ACM International Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*, pages 69–74, 2012.

I. Tabus, V. Tabus, and J. Astola. Information theoretic methods for aligning audio signals using chromagram representations. In *Proceedings of the 5th International Symposium on Communications Control and Signal Processing (ISCCSP)*, pages 1–4, 2012.

F. Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*, pages 366–381. Springer, 1981.

W. Tsai, H. Yu, and H. Wang. A query-by-example technique for retrieving cover versions of popular songs with similar melodies. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, pages 183–190, 2005.

J. W. Tukey. *The Problem of Multiple Comparisons*. Princeton University Press, 1973.

A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, Series 2*, 42:230–265, 1936.

A. Tversky. Features of similarity. *Psychological Review*, 84(4):327, 1977.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

Y. Vaizman, B. McFee, and G. Lanckriet. Codebook-based audio feature representation for music information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(10):1483–1493, 2014.

F. Vignoli and S. Pauws. A music retrieval system based on user driven similarity and its evaluation. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, pages 272–279, 2005.

V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

K. West, S. Cox, and P. Lamere. Incorporating machine-learning into music similarity estimation. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, pages 89–96, 2006.

R. P. Whorley, G. A. Wiggins, C. Rhodes, and M. T. Pearce. Multiple viewpoint systems: Time complexity and the construction of domains for complex musical viewpoints in the harmonization problem. *Journal of New Music Research*, 42(3):237–266, 2013.

G. A. Wiggins. Models of musical similarity. *Musicae Scientiae*, 11(1 suppl.):315–337, 2007.

G. A. Wiggins, D. Müllensiefen, and M. T. Pearce. On the non-existence of music: Why music theory is a figment of the imagination. *Musicae Scientiae*, 14(1):231–255, 2010.

J. Wülfing and M. Riedmiller. Unsupervised learning of local features for music classification. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 139–144, 2012.

Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H.-H. Chen. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):448–457, 2008.

J. E. Youngblood. Style as information. *Journal of Music Theory*, pages 24–35, 1958.

J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.

J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Glossary of Abbreviations

AIC algorithmic information content. 73–76, 80

ANN artificial neural network. 31, 32, 61, 63, 131, 137

ANOVA analysis of variance. 119, 125

BA balanced accuracy. 113, 115, 116, 118, 119, 123

BW Burrows-Wheeler. 48, 83–85, 91, 92, 95–97, 99, 101–103, 134, 136

CPE common practice era. 23–25

CPN common practice notation. 23, 25

DTW dynamic time warping. 45–49, 51, 135–137

ENR elastic net regularisation. 115–117, 122, 123

FCD feature complexity descriptor. 105–111, 115–128, 130–135, 138

FFN feed-forward network. 31, 33

FMD feature moment descriptor. 106, 108, 113, 116–128, 130–132, 138

GMM Gaussian mixture model. 56–60

HMM hidden Markov model. 44, 132, 133

I-R implication-realisation. 29, 30

IR information rate. 37, 38

JSD Jensen-Shannon divergence. 92, 95, 97, 99–101

KLD Kullback-Leibler divergence. 57–59, 62, 69, 71, 79, 92, 113, 116–119, 121

KNN K-nearest neighbours. 58, 121

LSH locality-sensitive hashing. 49, 50, 63

LSTM long short-term memory. 32, 133

LTM long-term model. 34

LZ Lempel-Ziv. 19, 34, 48, 60, 73, 79, 83–85, 90–92, 95–99, 101, 102, 110, 130, 133, 134, 136

MAE mean absolute error. 123–126

MAP mean average precision. 92–95, 97–101, 131, 137

MAR multivariate autoregressive. 55, 59

MFCC Mel-frequency cepstral coefficient. 36, 41, 43, 53–57, 60, 62, 108–110, 122, 126

MIR music information retrieval. 11, 12

MSD Million Song Dataset. 78, 89, 90, 94, 95, 97–102, 130, 131

MSE mean squared error. 88

NCD normalised compression distance. 17–19, 48, 49, 51, 60, 76, 78–80, 82–85, 91, 95–102, 129, 130, 133–135, 137

NCDA normalised compression distance with alignment. 19, 21, 81–85, 90, 91, 95–102, 129, 130, 133–137

NID normalised information distance. 75, 76, 78, 80, 81, 84, 86, 89, 92, 98–101, 134

NMSE normalised mean squared error. 88, 95, 98–101, 131, 132

OMSI Ollen musical sophistication index. 112

OTI optimal transposition index. 45, 90, 95

PBR preferred beat rate. 90

PCA principal components analysis. 41, 63, 110

PIR predictive information rate. 35, 37, 38

PPM prediction by partial matching. 33, 34, 48, 73, 83–85, 91, 92, 95–99, 101–103, 110, 134, 136

RBM restricted Boltzmann machine. 33, 61

RMSE root mean squared error. 123–125

RNN recurrent neural network. 31–33

STFT short-time Fourier transform. 41–43, 53

STM short-term model. 34, 36

SVM support vector machine. 58–60, 137

VMM variable-order Markov model. 33, 34, 36