Computational analysis of world music corpora

Maria Panteli

Submitted in partial fulfillment of the requirements
of the Degree of Doctor of Philosophy

School of Electronic Engineering and Computer Science
Queen Mary University of London

April 2018

I, Maria Panteli, 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:

Details of collaboration and publications: see Section 1.4.

Abstract

The comparison of world music cultures has been considered in musicological research since the end of the 19th century. Traditional methods from the field of comparative musicology typically involve the process of manual music annotation. While this provides expert knowledge, the manual input is time-consuming and limits the potential for large-scale research. This thesis considers computational methods for the analysis and comparison of world music cultures. In particular, Music Information Retrieval (MIR) tools are developed for processing sound recordings, and data mining methods are considered to study similarity relationships in world music corpora.

MIR tools have been widely used for the study of (mainly) Western music. The first part of this thesis focuses on assessing the suitability of audio descriptors for the study of similarity in world music corpora. An evaluation strategy is designed to capture challenges in the automatic processing of world music recordings and different state-of-the-art descriptors are assessed.

Following this evaluation, three approaches to audio feature extraction are considered, each addressing a different research question. First, a study of singing style similarity is presented. Singing is one of the most common forms of musical expression and it has played an important role in the oral transmission of world music. Hand-designed pitch descriptors are used to model aspects of the singing voice and clustering methods reveal singing style similarities in world music. Second, a study on music dissimilarity is performed. While musical exchange is evident in the history of world music it might be possible that some music cultures have resisted external musical influence. Low-level audio features are combined with machine learning methods to find music examples that stand out in a world music corpus, and geographical patterns are examined. The last study models music similarity using descriptors learned automatically with deep neural networks. It focuses on identifying music examples that appear to be similar in their audio content but share no (obvious) geographical or cultural links in their metadata. Unexpected similarities modelled in this way uncover possible hidden links between world music cultures.

This research investigates whether automatic computational analysis can uncover meaningful similarities between recordings of world music. Applications derive musicological insights from one of the largest world music corpora studied so far. Computational analysis as proposed in this thesis advances the state-of-the-art in the study of world music and expands the knowledge and understanding of musical exchange in the world.

Contents

List of Figures

List of Tables

Acknowledgements

This thesis wouldn't have been possible without the support of many people. First and foremost, I would like to thank my supervisor Simon Dixon for his excellent guidance over the years, his careful proofreading and critical questions, and for giving me freedom to explore and grow as a researcher. I am also very grateful to my second supervisor Emmanouil Benetos for his support of my work and inspiring conversations, and for always providing me with solutions for any research problem I discussed with him.

I am also grateful to Juan Pablo Bello for supporting my research visit at MARL-NYU, finding the time to supervise me, and encouraging me to collaborate with his group. I would also like to thank Matthias Mauch and Armand Leroi for helping me set the grounds in the first year of my PhD, and Mahey Mahendra for facilitating my research with the data from the British Library.

Many thanks also to the members of C4DM for the good times throughout the years, and the MARL and Spotify gangs for the enjoyable summers. Special thanks to Katerina, Rachel, Veronica, and Juanjo for all the good moments in conference travels and beyond, and to Jan for his endless support and patience. Many thanks finally to my family and friends for keeping me sane all this time!

This work was supported by a Queen Mary Principal's research studentship.

Chapter 1

Introduction

1.1 Motivation

The large-scale comparison of world music cultures has been the topic of several ethnomusicological studies since the end of the 19th century. Musicologist Guido Adler defined comparative musicology as part of systematic musicology, one of the two major subdisciplines of musicology (Adler, 1885). The term ethnomusicology was adopted to replace comparative musicology, but its concept is not only to study the world's musics from a comparative perspective but also to expand on the role of music within a culture and as a reflection of culture (see also Section 2.2). Comparative musicologists have made great progress in music data collection and analysis (Lomax, 1968, 1976; Brown and Jordania, 2011; Brown et al., 2014; Savage et al., 2015a). Though traditional forms of musicological analysis provide a great deal of expert knowledge, the manual annotation involved in the process is time-consuming and limits the potential for large-scale insights.

Today, the advances of technology in the field of Music Information Retrieval (MIR) (Schedl et al., 2014) allow for a thorough computational analysis of large music collections. The application of MIR techniques for the study of world music falls under the subdiscipline of Computational Ethnomusicology (Tzanetakis et al., 2007). Several research projects have focused on the development of MIR tools for the study of specific world music corpora (Marolt, 2009; Abdallah et al., 2017; Serrà et al., 2011; Fillon et al., 2014; Kroher et al., 2016). Applications of MIR tools to the study and comparison of large world music corpora however are yet to be explored.

Digital music archives today, such as the Smithsonian Folkways Recordings[1]

[1] `http://folkways.si.edu/folkways-recordings/smithsonian`, accessed 14th November 2017.

and the British Library Sound Archive[2], provide access to recorded material and associated metadata from various ethnomusicological collections. The aim of this research is to analyse world music corpora and explore relationships of music similarity and dissimilarity between different geographical areas of the world. The analysis focuses on processing music information from sound recordings using computational tools from the field of MIR. In particular, signal processing and data mining tools are developed to answer the following research questions.

What are the singing styles in world music? Singing is one of the most common forms of musical expression and the use of pitch by the singing voice (or other instruments) is recognised as a 'music universal', i.e., its concept is shared amongst all music cultures of the world (Brown and Jordania, 2011). Singing also plays an important role in the transmission of oral music traditions, especially in folk and traditional music styles. This research question focuses on modelling properties of the singing voice and comparing them to discriminate singing styles in world music and identify which music cultures share their singing characteristics.

Which music cultures are most distinct? The history of cultural exchange in the world goes back many years and music, an essential cultural identifier, has travelled beyond country borders. But is this true for all countries? What if a country is geographically isolated or its society resisted external musical influence? Can we find such music examples whose characteristics stand out from other musics in the world? Can we quantify the differences and relate them to musical attributes of rhythm, melody, timbre or harmony? This research question focuses on studying relationships of music dissimilarity in a world music corpus.

Which music cultures are unexpectedly similar? When considering world music we expect some music cultures to be more similar than others due to geographical and cultural proximity. For example, music from neighbouring countries or music performed from the same ethnic group and language might share some characteristics. But what if some music cultures appear to be similar but don't share any (obvious) geographical or cultural connections? Can we find such music examples that are *unexpectedly* similar? Can we explain this similarity by other factors? This research question incorporates prior knowledge of geographical and cultural aspects to model expectations of music similarity and investigates possible hidden links between world music cultures.

This research investigates whether automatic computational analysis can uncover meaningful similarities between recordings of world music, and if so what methods (audio features and data mining algorithms) are most suitable for each task. Computational tools aid the study of world music corpora and

[2] `http://sounds.bl.uk/World-and-traditional-music`, accessed 14th November 2017.

large-scale analysis increases the impact and certainty of the findings. The analysis of world music cultures as proposed in this study opens new directions for musicological research and expands the knowledge and understanding of musical exchange in the world.

1.2 Contributions

The contributions of this thesis are manifold:

- A corpus of music data from around 60000 sound recordings and associated metadata is curated for the purposes of a computational world music comparison (Chapter 3). The audio is not available due to copyright protection[3]. Following signal processing techniques adopted in this research, the audio signal is represented in the (non-reversible) form of Mel-scaled (Stevens et al., 1937) spectrograms. Sharing this corpus will facilitate the continuation of research in world music with computational methods such as the ones described in this thesis.

- The second contribution is the assessment and adaptation of existing audio features for the large-scale processing of world music recordings. Chapter 4 discusses the challenges of audio feature extraction in world music and proposes an evaluation strategy to assess the suitability of audio features for world music similarity. Part of the contribution is the conclusions, i.e. which features are found to be best.

- In MIR literature, models of music similarity have been developed for mainly Western music. This thesis contributes techniques for modelling music similarity in world music corpora. In particular, three different approaches are followed in this thesis: a) the custom feature design of pitch descriptors (Chapter 5), b) the combination of low-level music descriptors with linear transformation methods to capture aspects of music style (Chapter 6), and c) learning of high-level music features directly from world music data with deep learning methods (Chapter 7). These approaches expand the state-of-the-art of MIR research in world music.

- The application of computational models to test musicological hypotheses provides valuable insights in the history of musical exchange in the world. The analyses in Chapters 5, 6, and 7 reveal geographical patterns of music similarity. These are the first findings from a large-scale world music

[3]Short audio previews can be accessed through the official websites of the Smithsonian Folkways Recordings (`http://www.folkways.si.edu/folkways-recordings/smithsonian`) and British Library Sound Archive (`http://sounds.bl.uk/World-and-traditional-music`).

comparison with computational tools. There is a lot to be explored yet and this research forms a basis for future computational studies in world music.

1.3 Thesis Outline

The remainder of this manuscript is organised as follows:

- **Chapter 2** defines the terminology, and reviews state of the art approaches to music corpus analysis with the focus on world music. It also discusses criticism received by major comparative studies, presents the challenges associated with a computational world music comparison such as the one considered in this thesis, and concludes with a discussion of how state of the art can be improved.

- **Chapter 3** presents the world music corpus used for this research with descriptive statistics derived from the metadata of the sound recordings. It also describes the semi-automatic techniques used for curating the metadata and the way additional corpora were derived from this large world music corpus to assess musicological hypotheses and evaluate computational algorithms.

- **Chapter 4** describes the main methodology of this thesis which relies on the extraction of audio features from the spectrogram. It reviews different techniques for audio feature extraction as used in the literature so far, discusses the applicability of these methods for the analysis of world music corpora, and presents an evaluation strategy to assess the performance of audio features for the study of world music similarity.

- **Chapter 5** presents the development of a singing similarity model and its application to compare the singing styles in world music. In this chapter, specifically designed audio features model particularities of the singing voice and dictionary learning methods are applied to extract the singing vocabulary in world music. Analyses focus on clustering singing styles in world music and exploring singing style similarities between different music cultures.

- **Chapter 6** presents the development of a music dissimilarity model and its application to identify music examples that are most distinct in the corpus. It presents a method to learn a feature space for music similarity combining low-level audio features with machine learning methods and a way to quantify music dissimilarity based on outlier detection techniques.

Various analyses explain the characteristics of distinct music examples and reveal geographical patterns of music outliers in the world.

- **Chapter 7** presents the development of a music similarity model and its application to identify music examples that appear to be unexpectedly similar. Music similarity expectations are defined based on geographical and cultural links derived from the recordings' metadata. In contrast to Chapter 6, a music similarity model is learned via the application of deep learning methods. The analyses focus on explaining music similarity relationships and studying geographical patterns of unexpectedly similar music.

- **Chapter 8** discusses directions of improvement for future work and summarises the findings and contributions of this thesis.

1.4 Publications

The main publications associated with this thesis are the following:

1. Panteli, M., & Dixon, S. (2016). On the evaluation of rhythmic and melodic descriptors for music similarity. In *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 468-474).

2. Panteli, M., Benetos, E., & Dixon, S. (2016). Learning a feature space for similarity in world music. In *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 538-544).

3. Panteli, M., Benetos, E., & Dixon, S. (2016). Automatic detection of outliers in world music collections. In *Analytical Approaches to World Music Conference* (pp. 1-4).

4. Panteli, M., Benetos, E., & Dixon, S. (2017). A computational study on outliers in world music. *PLOS ONE*, 12(12): e0189399. `http://doi.org/10.1371/journal.pone.0189399`.

5. Panteli, M., Benetos, E., & Dixon, S. (2018). A review of manual and computational approaches for the study of world music corpora. *Journal of New Music Research*, 47:2, 176-189, DOI:10.1080/09298215.2017.1418896.

Other publications where the author contributed to:

6. Panteli, M., Bittner, R., Bello, J. P., & Dixon, S. (2017). Towards the characterization of singing styles in world music. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 636-640).

7. Kedyte, V., Panteli, M., Weyde, T., & Dixon, S. (2017). Geographical origin prediction of folk music recordings from the United Kingdom. In *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 664-670).

In (6) the author curated the world music dataset, developed code to learn a dictionary of (pre-computed) singing style elements, performed the singing clustering and analysed the results. In (7) the author curated the dataset, suggested the design of MIR features from the output of VAMP plugins, and partly analysed the results. In all other cases the author was the main contributor to the publication, under supervision of Dr. Emmanouil Benetos and/or Prof. Simon Dixon.

The above publications relate to the chapters of this thesis in the following way. Publication (5) reviews the state-of-the-art and is the basis of Chapter 2 and publication (1) considers the evaluation of audio features and is the basis of Chapter 4 and more specifically Section 4.4. Publication (6) is the basis of the singing style similarity study (Chapter 5). Publications (2) and (3) discuss the development of the methodology for the music dissimilarity model (Chapter 6) and publication (4) presents the results from the application of this model. Publication (7) presents another approach to feature learning from world music data and relates to the developments in Chapters 4 and 7.

Chapter 2

Related work

This chapter reviews related work on the analysis of world music corpora. It focuses on music corpus studies from the fields of ethnomusicology and MIR with manual or computational approaches. This chapter defines also the terminology used throughout this thesis. Beyond the literature review, a summary of opposing views for the comparison of world (or other) music styles is presented. Lastly, the challenges involved in the computational study of world music corpora are explained, some of which are tackled in the developments of this thesis (e.g., the choices of sampling world music corpora as presented in Chapter 3 and audio processing methods as presented in Chapter 4).

2.1 Background

The fields of ethnomusicology and MIR have set the grounds for a large-scale comparison of world music. These fields bring different expertise to the challenging study of world music and the collaboration between the two has been considered a great advantage (van Kranenburg et al., 2010; Neubarth et al., 2011; Inskip and Wiering, 2015). In this thesis large-scale computational analysis of world music integrating knowledge from both ethnomusicology and MIR is considered. Related work reviewed here covers a comparison of the size and scope of music corpora used in manual and computational approaches, the research questions and findings, as well as the music descriptors and data mining tools used in each study. Major comparative studies have also received criticism (Nettl, 1970; Clarke, 2014; Fink, 2013; Underwood, 2015). This chapter highlights the strengths and weaknesses of state of the art research and points toward remaining challenges and lessons learnt for an improved computational study of music.

There are numerous manual and computational approaches for the compar-

ison of world music cultures. Studies reviewed here are selected based on four primary criteria. The first two criteria are a) the size of the corpus they analyse and b) the research question they address. In particular, I exclude computational studies whose research question is not targeted at understanding the corpus itself but rather at assessing the methods or pursuing a specific algorithmic challenge. Likewise, I exclude manual studies that explore a relatively small (less than 100 recordings) or very specific corpus as it is unlikely that the methods are scalable and generalisable to non-specific corpora. The other two criteria require that c) the studies under review are primarily concerned with the comparison of music cultures and d) they provide a rather systematic approach in their methodology. The primary interest of this thesis is the comparison of world music cultures but since not many studies have considered a world music corpus this review is expanded to include comparative music studies on popular, classical, and folk and traditional music repertoires. The review is primarily focused on studies that process music information from the sound recording or the music notation. World music studies based on historical, cultural, or other metadata information (Barrett, 1996; Baily and Collyer, 2006) are beyond the scope of this thesis.

2.2 Terminology

Terms and concepts frequently used in this thesis are explained in the paragraphs below.

One of the most ambiguous terms is that of *world music*. The term can have various interpretations, and throughout the literature it has been used to denote popular and classical musics from around the world and from different eras (Bohlman, 2002). *World music* in this thesis is used to define folk and traditional music from around the world, including Western folk music but excluding Western art music. Another controversial term is that of *folk music*. According to Nettl (2014), "folk music is a type of traditional and generally rural music that originally was passed down through families and other small social groups. Typically, folk music lives in oral tradition; it is learned through hearing rather than reading. It is functional in the sense that it is associated with other activities, and it is primarily rural in origin." *Music corpus* defines a collection of music pieces in recorded form or musical notation. A *corpus-based study* addresses primarily research questions regarding the characteristics of the music corpus.

The developments in the thesis often refer to the study and comparison of music styles or music cultures. *Music style* refers to "characteristic uses of form, texture, harmony, melody, and rhythm" (Sadie et al., 2001). *Music culture*

defines the music performed by a specific ethnic group, society or geographical region. *World music cultures* denote the different musics performed around the world. Since the medium of analysis throughout the thesis is the sound recording, *culture* denotes characteristic musical aspects captured solely by the sound recording. The state-of-the-art review presented in this chapter covers both studies with sound recordings as well as music notation. The terms music style and music culture are sometimes used interchangeably in this thesis.

References to two major research fields, ethnomusicology and MIR, are frequently made. *Ethnomusicology* traditionally focused on the study of non-Western music of oral traditions but today expands to the study of all music of the world in its social and cultural context (Pegg et al., 2001; Dahlig-Turek et al., 2012). The term ethnomusicology was adopted to replace *comparative musicology* (Adler, 1885), but its concept is not only to study the world's musics from a comparative perspective but also to expand on the role of music within a culture and as a reflection of culture (Nettl, 2005). Another related field is *systematic musicology* (Adler, 1885), which includes the study of collections of music using analytical, statistical, or computational approaches (Leman, 2008). In contrast to other fields of musicology, systematic musicology addresses "how music practices can be understood, explained as a system (both from a psychoneuronal and social point of view), and possibly further explored and exploited (for example in connection with technology)" (Leman, 2008, p. 1).

MIR is foremost concerned with the extraction and inference of musically relevant features (from the audio signal, symbolic representation or external sources such as web pages), indexing of music using these features, and the development of different search and retrieval schemes (for instance, content-based search, music recommendation systems, or user interfaces for browsing large music collections) (Schedl et al., 2014). *Digital musicology* is defined as interdisciplinary music research which encourages the use of technical infrastructure for musicology (Wiering and Benetos, 2013), often referred to also as *computational musicology* (Bel and Vecchione, 1993; Volk et al., 2011). *Computational ethnomusicology* refers to the application of computational algorithms for the study of, specifically, non-Western music corpora (Tzanetakis et al., 2007; Gómez et al., 2013).

I also make the following distinctions. The medium of music representation studied in the various manual and computational approaches reviewed in this chapter can be either the *sound recording* or *music notation*. The former captures an acoustic representation of music as an audio signal whereas the latter defines a symbolic representation of music as a score or other music notation system.

The systematic description of music can be made with either *manual anno-*

tations or *automatically extracted features*. The former denotes the process of human experts manually annotating musical attributes for each recording, for example the Cantometrics and Cantocore systems for world music (Lomax, 1976; Savage et al., 2012) and the Music Genome project for Western popular music (Prockup et al., 2015). Automatic feature extraction denotes the computational approach to derive musical attributes from the audio signal, for example using audio processing (McFee et al., 2015b; Tzanetakis and Cook, 2000; Lartillot and Toiviainen, 2007) or music notation processing (McKay, 2010) software. I refer to studies based on human annotations to music description as *manual approaches* and studies based on automatically extracted features as *computational approaches*. Manual approaches could still employ computational methods at a later stage of the analysis. However, the initial music annotation (human or computational) that the analysis is based on, is what defines the approach as manual or computational throughout this chapter.

2.3 Music corpus studies

2.3.1 Manual approaches

Many studies in the field of ethnomusicology have considered and discussed the comparison of music cultures (Feld, 1984; Tenzer, 2006; Nettl and Bohlman, 1991; Nettl, 2015). Feld (1984) reflects on the approaches of comparative music studies and discusses the need for a qualitative comparison as well as the research questions that could contribute to the understanding of socio-musical practices. Tenzer (2006) explores music repertoires from around the world and reviews the contexts of their performance and creation and the ways to hear and conceive the different musical attributes. Nettl and Bohlman (1991) discuss the methodological and theoretical foundations as well as significant issues in the history of ethnomusicology. Nettl (2015) provides an overview of ethnomusicological research and focuses on concepts and issues that have caused a long ethnomusicological discourse.

A review of comparative music studies is also presented by Savage and Brown (2013). The authors redefine the field of comparative musicology, revisiting the research goals and discussing potential contributions of the field to the study of music classification, cultural evolution, human history, music universals, and biological evolution. In this chapter I review a subset of these studies matching the criteria defined in the first paragraph of Section 2.1 and expand on music studies with computational approaches, as the main focus of this thesis.

Audio recordings

One of the major comparative musicologists in the 1960s was Alan Lomax who collected more than 4000 recordings from many geographical areas and developed an annotation system, *Cantometrics* (Lomax, 1976), to categorise the music cultures of the world (Lomax, 1968). Using a phylogenetic analysis, Lomax (1980) suggested the existence of two evolutionary roots, the Siberian and African Gatherers music styles (Lomax, 1980, p. 39). More recently, Savage and Brown (2014) analysed 259 traditional songs from 12 indigenous populations from Taiwan using 26 features from the *Cantocore* system (Savage et al., 2012) focusing on rhythm, pitch, texture, and form. Using clustering analysis Savage and Brown (2014) showed that songs can be grouped in 5 clusters correlated with geographical factors and repertoire diversity. With a smaller corpus of 72 songs, Mora et al. (2010) developed a set of manual annotations for two flamenco styles, deblas and martinetes, and measured inter- and intra-style similarity with Euclidean distances and phylogenetic trees. A related study (Kroher et al., 2014) investigated similarity measures based on manually annotated and computationally extracted flamenco features and compared these measures to human ratings of melodic similarity.

Another application of comparative musicology is in the search for musical universals, i.e. the systematic comparison of the world's musics in order to understand which features are universal and which are culture-specific. Brown and Jordania (2011) proposed 70 music features to model aspects of the music structure and performance, as well as social features related to the social context, content, and music-related behaviours. Based on qualitative analysis the authors proposed a set of features, such as the use of musical scales, octave equivalence, amplitude and tempo dynamics, and music organisation into phrases, that describe the properties of (most) musical systems throughout the world. Expanding on the topic of music universals, Savage et al. (2015a) analysed 304 recordings contained in the Garland Encyclopedia of World Music (Nettl et al., 1998) using 32 features from the Cantocore and Cantometrics systems and instrument classification attributes as defined by von Hornbostel and Sachs (1961). Using phylogenies to control for historical relationships, continuous Markov processes to model rate of change and correlations of features across cultures, they were able to show that there are no *absolute* music universals but rather *statistical* universals. For example, there are 18 music features shared amongst many music cultures of the world and a network of 10 features that often occur together.

Other music comparative studies have focused on contrasting music to genetic and language evolution. Rzeszutek et al. (2012) annotated 421 traditional

songs from 16 Austronesian-speaking populations from Taiwan and the northern Philippines using the Cantocore system. Correlations between music and genes showed that the majority of musical variability is due to differences within populations rather than differences between populations. In a similar study with 220 traditional songs from 9 indigenous populations from Taiwan, and a set of 41 descriptors (26 from Cantocore and 15 from Cantometrics systems), Brown et al. (2014) showed that population structures for genetics indicate stronger parallels to music than to language. Savage et al. (2015b) compared genetic and musical diversity by analysing 680 traditional songs from two Ainu and 33 East Asian and circumpolar populations. The distribution of stylistic song-types in music was similar to the distribution of DNA types and consistent with a 'triple structure' model of Japanese archipelago history. Le Bomin et al. (2016) analysed 700 recordings from 58 patrimonies of rural areas in Gabon using 322 features on repertoire, form, instrument, metre, rhythm, and melody. A phylogenetic analysis of repertoires showed that there is a predominant vertical transmission of musical characteristics such as metre, rhythm, and melody, where vertical transmission refers to the inheritance from ancestors in contrast to horizontal exchange from neighbours.

Music notation

A few studies were found using manual approaches to explore relatively large corpora of music notation. Bronson (1950) analysed several melodic, rhythmic, and structural attributes of 100 British folk tunes from the 16th to the 20th century. His findings include comparative statistics of the use of tune length, modes, metres, cadences, and phrase patterns over the time span of five centuries. A related study (Savage, 2017), analysed 4125 British-American narrative songs from the Child ballads collection (Bronson, 1972) notated between $1575 - 1972$. Hypotheses related to music culture evolution were tested and analysis showed that, amongst others, "functional notes are more resistant to change than ornamental notes and substitutions are more likely to occur between small melodic distances than large ones" (Savage, 2017, p. 68). Volk and van Kranenburg (2012) developed an annotation method for 360 Dutch folk melodies including features capturing aspects of contour, rhythm, and motif similarity. They found that the recurrence of characteristic motifs is the most important feature for establishing similarity in Dutch folk melodies. Musicological hypotheses were also tested in a study of harmonic usage in American popular music as it evolved from the 1950s to the 1990s (Burgoyne et al., 2013). The authors used 1379 songs from the Billboard dataset with chord transcriptions manually annotated by experts (Burgoyne et al., 2011), and performed compositional data analysis

to illustrate changes in harmonic usage over time. They found that there is a greater use of minor tonalities over time and dominant chords become less frequent than tonic and subdominant chords in recent songs.

A number of studies that have explored statistical techniques for the analysis of specific music notation corpora can be reviewed in Nettheim (1997), Temperley and Van Handel (2013), Gustar (2014), Walshaw and Walshaw (2014) and references therein. The majority of these studies focus on either small corpora or corpora and methods of very specific music styles and are thus beyond the scope of this review. It is also worth noting here that many world music cultures are orally transmitted and the resources of music notation are often limited. What is more, the study of music notation corpora employed computational tools from an early stage (Bronson, 1949; Scherrer and Scherrer, 1971) and therefore these are summarised under computational approaches in Section 2.3.2.

A summary of the manual approaches reviewed above is shown in Table 2.1.

2.3.2 Computational approaches

The use of computers for the comparison or classification of music cultures has been considered as early as the middle of the 20th century (Bronson, 1949; Rhodes, 1965). Music corpus studies using computational tools have been considered in the fields of MIR and computational musicology. In these studies the corpus is usually larger due to the efficiency of computational analysis but questions are raised on how representative and meaningful the automatically extracted features are. Below I review computational approaches using sound recordings and music notation. These approaches are also the main source and inspiration for the world music analysis methods developed in this thesis.

Audio recordings

A number of computational approaches have focused on studying stylistic characteristics as they evolve over time. A study of 1010 recordings from the top 40 of the Billboard Hot 100 charts between $1965-2009$ revealed that popular recordings became longer in duration and more sad-sounding over time (Schellenberg and von Scheve, 2012). Serrà et al. (2012) analysed pitch, timbre, and loudness in 464411 recordings (between $1955-2010$) of Western popular genres from the Million Song Dataset (MSD) (Bertin-Mahieux et al., 2011). Analysing music trends over the years revealed that more recent music shows less variety in pitch transitions, consistent homogenisation of the timbral palette, and louder and potentially poorer volume dynamics. Shalit et al. (2013) used 24941 songs by 9222 artists (between $1922-2010$) from the Million Song Dataset, audio features related to pitch, timbre, and loudness, and topic models (Blei and Lafferty,

Study	Corpus size and description		Descriptor count and type		Main Findings
Manual approaches - Audio recordings					
(Lomax, 1968)	1800	Various world music cultures	36	Cantometrics	Two possible evolutionary roots, the Siberian and African Gatherers music styles.
(Rzeszutek et al., 2012)	421	Traditional music from Taiwan and Philippines	26	Cantocore	The majority of musical variability is due to differences within populations rather than between populations.
(Savage and Brown, 2014)	259	Traditional music from Taiwan	26	Cantocore	Songs grouped in 5 clusters correlated with geographical factors and repertoire diversity.
(Savage et al., 2015a)	304	World music from the Garland encyclopedia	32	Cantocore and Cantometrics	No *absolute* music universals but *statistical* universals (18 common features and a network of 10 co-occurring features).
(Savage et al., 2015b)	680	Traditional songs from Japan and East Asia	41	Cantocore and Cantometrics	Stylistic song-types in music have similar distribution to DNA types and are consistent with a 'triple structure' model.
(Le Bomin et al., 2016)	700	Traditional music from Gabon	322	Form, instrument, rhythm, melody	Predominant vertical transmission of musical characteristics such as metre, rhythm, and melody.
Manual approaches - Music notation					
(Bronson, 1950)	100	British folk tunes	-	Mode, meter, form, cadence	Comparative statistics of the use of musical attributes over 5 centuries.
(Volk and van Kranenburg, 2012)	360	Dutch folk songs	-	Melodic contour, rhythm, motif	The recurrence of characteristic motifs is the most important feature for establishing melodic similarity in Dutch folk melodies.
(Burgoyne et al., 2013)	1379	American popular music from the Billboard set	-	Beat-aligned chord transcriptions	There is a greater use of minor tonalities over time and dominant chords are less frequent than tonic and subdominant chords.
(Savage, 2017)	4125	British-American narrative songs	-	Melodic contours	Functional notes are more resistant to change than ornamental notes and substitutions are more likely to occur between small melodic distances than large ones.

Table 2.1: The size of corpus, type of features, and findings of music corpus studies in the literature using manual approaches and subject to the selection criteria defined in Section 2.1. The count of descriptors is denoted by '-' when the exact number of features was not explicitly stated in the corresponding published work.

2006), and showed that the most influential songs were more innovative during the early 1970's and the mid 1990's than at other times. Mauch et al. (2015b) studied harmonic and timbral content in 17094 songs covering 86% of the US Billboard Hot 100 between $1960 - 2010$. Using topic modelling and clustering analysis they concluded that USA pop music evolved with particular rapidity during three stylistic 'revolutions' around 1964, 1983 and 1991.

With respect to non-Western music repertoires, Moelants et al. (2009) studied pitch distributions in 901 recordings from Central Africa[1]. They observed that music from Central Africa does not conform to the 12-tone equal temperament, however in recent recordings there seems to be a tendency to the use of more equally-tempered scales. Gómez et al. (2009) studied music style classification in a dataset of 5905 recordings of Western and non-Western traditions using tonal, timbral, and rhythmic features. Their analysis verifies that Western music is more equal-tempered than non-Western and an investigation of which features correlate most with geographical regions indicated that latitude is mostly associated with tonal features and longitude with rhythmic ones. Other approaches to non-Western music analysis include the automatic classification of audio recordings into global cultural areas (Kruspe et al., 2011; Zhou et al., 2014), classification of ethnomusicological recordings by timbre features (Fourer et al., 2014), the study of pitch distributions in Turkish (Bozkurt, 2008), Byzantine (Panteli and Purwins, 2013), and Indian classical (Ganguli et al., 2016) music, rhythmic patterns in Turkish (Holzapfel and Stylianou, 2009) and Indian art (Srinivasamurthy et al., 2014) music, and the development of computational models for investigating similarity in world music corpora (Holzapfel, 2010; Panteli et al., 2016a).

Music notation

Computational approaches have also been applied to analyse music in symbolic representation. A study of melodic contours from 6251 European folk songs from the Essen Folksong Collection (Schaffrath, 1995) revealed that melodies tend to exhibit an arc-shaped pitch contour (Huron, 1996). Zivic et al. (2013) analysed classical music scores between $1700 - 1930$ from the Peachnote corpus (Viro, 2011) which consists of more than 900000 scores. By studying bigrams of melodic intervals they were able to show that classical music styles are distinguished by characteristic differences in their distribution of melodic intervals over time. Pamjav et al. (2012) analysed pitch sequences of 31 Eurasian and North-American folksong collections, each of them consisting of $1000 - 2500$ melodies. Using Self Organising Maps (SOMs) and Multi-Dimensional Scaling

[1]The Royal Museum for Central Africa `http://music.africamuseum.be`, accessed 14th November 2017.

approaches they showed that there is a significant correlation between population genetics and folk music, and that maternal lineages in folk music are more prominent than paternal lineages. Volk and de Haas (2013) studied syncopation in ragtime music by analysing melodic patterns from 11000 ragtime MIDI files. The authors confirmed the musicological hypothesis that the use of tied syncopations increased in the ragtime era after 1902 in comparison to the use of untied syncopations.

Aarden and Huron (2001) analysed the phrase endings from European folk melodies of the Essen Folksong Collection. From a total of approximately 950 melodies they observed that Western European melodies are more likely to have their melodies ending on the tonic than Eastern European melodies. Juhász (2006) studied melodic contours of approximately 9000 folksongs from Slovak, French, Sicilian, Bulgarian, English, and Hungarian music cultures. Using SOMs it was shown that a common set of contour types was shared amongst the 6 cultures and that these contour types are represented especially in the Hungarian and Slovak traditions. In a subsequent study including music from additional cultures of Eurasia, SOMs analysis revealed that the use of melodic contours in different geographical areas can be grouped into two main clusters (Juhász, 2009). Shanahan et al. (2016) analysed 2083 folksongs from the Frances Densmore's collection of Native American music using attributes from the jSymbolic set (McKay, 2010) and information-theoretic measures. Contrast mining methods (Dong and Li, 1999) were employed to compare music in different social contexts. Their analysis showed, amongst others, that nature songs have low variability of events, love songs have larger melodic intervals and higher pitch registers, and war and dance songs are "high arousal" songs but on opposite ends of the valence spectrum on Russell's Circumplex model (Russell, 2003). Other approaches to studying music corpora include the classification of folk Dutch melodies with local and global features (van Kranenburg et al., 2013), and the analysis of melodic patterns in Cretan folk songs (Conklin and Anagnostopoulou, 2011).

A summary of the computational approaches reviewed above is shown in Table 2.2.

2.4 Criticism

Music corpus studies seen in the literature so far have received considerable criticism. In this section I review issues raised about the most popular comparative studies.

The work by Lomax (1976) has concerned ethnomusicologists and anthropologists (Dubinskas, 1983; Nettl, 1970; Feld, 1984). Some of the critical remarks as

Study	Corpus size and description		Descriptor count and type		Main Findings
Computational approaches - Audio recordings					
(Moelants et al., 2009)	901	Traditional music from Central Africa	1200	Pitch histogram	African music does not conform to the fixed chromatic scale but recent recordings use more elaborate, equally-tempered scales.
(Gómez et al., 2009)	5905	Traditional music from Western and non-Western countries	23	Timbre, rhythm, tonality	Western music is more equal-tempered than non-Western, latitude mostly associated with tonal features and longitude with rhythmic ones.
(Schellenberg and von Scheve, 2012)	1010	American popular music from the Billboard set	4	Tempo, mode, duration, gender of vocalist	Popular recordings became longer in duration and more sad over time.
(Serrà et al., 2012)	464411	Western popular music from the Million Song Dataset	-	Pitch, timbre, loudness	Recent music shows less variety in pitch transitions, consistent homogenisation of the timbral palette, and louder and potentially poorer volume dynamics.
(Shalit et al., 2013)	24941	Western popular music from the Million Song Dataset	-	Pitch, timbre, loudness	The most influential songs were more innovative during the early 1970's and the mid 1990's.
(Mauch et al., 2015b)	17094	American popular music from the Billboard set	-	Tonal and timbral topics	Pop music evolved with particular rapidity during three stylistic revolutions around 1964, 1983 and 1991.
Computational approaches - Music notation					
(Huron, 1996)	6251	European folk music from the Essen Collection	-	Melodic contours	European folk melodies tend to exhibit an arch-shaped pitch contour.
(Aarden and Huron, 2001)	950	European folk music from the Essen Collection	-	Melodic phrase ending	Western European melodies are more likely to have their melodies ending on the tonic than Eastern European melodies.
(Juhász, 2006)	9000	Folk music from 6 European cultures	-	Melodic contours	There is a common set of contour types shared amongst the 6 cultures and these contour types are represented especially in the Hungarian and Slovak traditions.
(Juhász, 2009)	19733	Folk music from 11 Eurasian cultures	-	Melodic contours	The use of melodic contours in different geographical areas can be grouped into two main clusters.
(Pamjav et al., 2012)	$\approx$ 31000	folk music from Eurasia and North-America	-	Melodic contours	There is a significant correlation between population genetics and folk music and maternal lineages in folk music are more prominent than paternal lineages.
(Zivic et al., 2013)	$\approx$ 900000	Western classical music	-	Melodic interval bigrams	Classical music styles show characteristic differences in their distributions of melodic intervals over time.
(Volk and de Haas, 2013)	11000	Ragtime music	-	Syncopation in melodic patterns	The use of tied syncopations in comparison to untied syncopations increased in the ragtime era after 1902.
(Shanahan et al., 2016)	2083	Native American folk music	18	Pitch, duration, intervals, onsets	Contrasts between musics in different social contexts, e.g., nature songs have low variability of events, love songs have larger melodic intervals.

Table 2.2: The size of corpus, the type of features, and the findings of music corpus studies in the literature using computational approaches and subject to the selection criteria defined in Section 2.1. The count of descriptors is denoted by '-' when the exact number of features was not explicitly stated in the corresponding published work.

Nettl (1970) suggests are that the dataset samples too few songs from each culture and that the annotation system (Cantometrics) may not be representative because the annotators may lack a complete understanding of the music: "Can someone understand a music without immersing oneself in it for years?" (Nettl, 1970, p. 439). Furthermore, annotations may not be very reliable due to the difficulty of the task for human listeners, "evaluating by ear such elusive qualities as vocal rasp, nasality, and vocal width (which are not standard or widely used concepts in musicology) and assigning their relative degree in a recording according to a scale of up to ten points would appear to be a questionable procedure" (Nettl, 1970, p. 440). Feld (1984) discusses the need for a qualitative and intensive comparative musicology and comments that "the best way to answer Lomax's questions about the systematic nature of musical representation in social organisation is to study them on the ground, in the field, up close, over long periods of time, where sound structures are observably and undeniably socially structured". He also defines research questions under six domains (competence, form, performance, environment, value and equality, theory) that could contribute to the comparison of socio-musical realities and practices.

Savage and Brown (2014) described key themes in comparative musicology and included, amongst others, the generation of a musical map of the world reflecting aspects of cultural diversity and evolution. Clarke (2014) criticises the properties of the music to be considered in the creation of such a map, "Should it be based on musical production (composition, performance), or consumption (concert going, private listening)? Should we consider the public sphere (larger, widely advertised events) or the (semi-)private (domestic get-togethers and community gatherings)?" (Clarke, 2014, p. 9). He also raises a point about temporal evolution, "Traditions evolve, styles mutate, patterns of consumption change", that is not captured in a static collection of music and a projected local map "would be just one snapshot on a much larger diachronic continuum" (Clarke, 2014, p. 9).

Large-scale computational approaches to music corpus analysis have also received criticism. One of the major issues for the study of Serrà et al. (2012) is the suitability of the corpus. Fink (2013) observes that the study investigates evolutionary trends in the Million Song Dataset, a dataset created primarily for the evaluation of MIR algorithms. As Fink (2013) mentions, "any conclusions drawn from the MSD are already constrained by the assumptions and mindset of the industry-research teams that created the database". Another major drawback is that the music coding system is not easily interpretable and numerical representations derived from the model can be questioned as to whether they contain meaningful musical information (Fink, 2013; Wallmark, 2013). What is more, Western bias may influence the interpretation of results (Fink, 2013) and

the social context in which the music is actually heard is disregarded in such computational analysis (Wallmark, 2013).

Similar critical remarks apply to the study by Mauch et al. (2015b). Underwood (2015) discusses whether measures of stylistic "distance" between songs can indicate cultural change and how robust these measures can be. In another post, Underwood et al. (2016) suggest that statistical significance is calculated in a misleading way, "only two of the three "revolutions" it [(Mauch et al., 2015b)] reported are really significant at $p < 0.05$, and it misses some odd periods of stasis that are just as significant as the periods of acceleration". Thompson (2015) points to some alternative factors, namely the change in the Billboard measurement system in 1991, that might have contributed to observing a music revolution in 1991 as concluded by Mauch et al. (2015b).

The critical remarks presented above for different studies in the literature are often overlapping. For example, the suitability of the corpus has been questioned in both manual (Lomax, 1976) and computational (Serrà et al., 2012) approaches. The reliability of music annotations can be an issue in both approaches (see for example remarks by Nettl (1970) and Fink (2013) above). On one hand human experts may not be able to reliably annotate fine-grained musical characteristics judging solely by ear. On the other hand computational systems may fail to capture high-level attributes for example aspects of music perception and cognition. The above criticism gives valuable feedback on challenges that need to be addressed for an improved world music study.

2.5 Challenges

In the literature reviewed above manual approaches have used relatively small corpora of world music (Lomax, 1980; Savage et al., 2015a), and computational approaches have used large corpora but focusing mainly on Western music (Serrà et al., 2012; Mauch et al., 2015b). A large-scale comparative study of world music cultures has not been addressed yet. Computational tools could aid such comparisons and large-scale analysis could increase the impact of any findings. However, a large-scale comparison with computational tools includes several challenges with respect to processing information from the metadata and the audio recordings as well as generalising findings from big data collections. Below I list the major challenges associated with this line of research.

Restricted access to audio recordings. While several research projects and institutions make great efforts to increase the accessibility to audio music collections (Porter et al., 2013; Franzen, 2016; Abdallah et al., 2017), a lot of recorded world music is still not available for research due to copyright and other

ethical issues[2]. To create a world music corpus it is necessary to combine sound recordings from distributed sources and collections. This brings up further challenges in setting up legal agreements with the owners of each collection and processing the information from each source in a unified manner. The Digital Music Lab project (Abdallah et al., 2017) proposed to circumvent this problem by performing the analysis locally on each collection and aggregating the results centrally.

Unbalanced collections. Access to fieldwork in ethnomusicology as well as in other ethnographic research is affected by spatial and temporal parameters (Hammersley, 2006; Barz and Cooley, 2008). In large collections of world music recordings it is often the case that Western-influenced music traditions are more represented than non-Western. A comparative study on world music however requires a balanced corpus with a good representation of the geographical and cultural diversity of world music as well as a good temporal spread of the music eras.

Corpus creation. Creating a corpus suitable for the computational study of world music imposes further challenges in terms of qualitative and quantitative criteria. As seen in past criticism (Section 2.4), the corpus needs to include the most representative samples from each music culture (Nettl, 1970), and the assumptions made to create the corpus must be in line with the research questions under study (Fink, 2013). This requires addressing what defines a good sample, how to balance the diversity, and how to maximise the size of this corpus to obtain large-scale results. Serra (2014) defines five criteria, namely the purpose, coverage, completeness, quality, and reusability to be taken into account when creating corpora for the computational study of music. Similar criteria are also followed by Kroher et al. (2016) for the creation of a corpus for the computational study of flamenco music.

Interpretation of metadata. In order to study the relationships between musical content and metadata of world music, spatio-temporal information of the origins of the music is required. In world music recordings the temporal information associated with the metadata represents the time the music was recorded but does not necessarily represent the time at which it was composed. For example, for most folk music the time and location of a song's composition remains unknown. What is more, unlike Western popular music where there is often a common agreement concerning the taxonomy of music styles, in world music the classification of music styles is still in great discourse (Lomax and Berkowitz, 1972; Clayton et al., 2003). The assumptions made when creating the

[2]In some cases, copyright exceptions encourage research with audio recordings as long as the research is non-commercial, the resources are properly acknowledged, and the research results cannot recreate the original works (see for example regulations for research in the UK at `http://www.gov.uk/government/organisations/intellectual-property-office`).

metadata need also to be considered, for example, the purpose of the metadata creation and the background and interest of the curators. There are therefore greater challenges involved in processing the metadata for world music.

Incorrect metadata. Depending on the collector and the era in which a recording session took place, the information registered for each recording varies vastly or is absent altogether. A great challenge is therefore to combine all the available information and create a consistent database of metadata. In several cases information on the culture or language of a recording is misspelt or the registered location is inconsistent with the latest geopolitical maps (e.g., USSR or Yugoslavia whose borders and political status have changed). Automatic correction of this type of metadata requires techniques from natural language processing and geopolitical database matching.

Lack of ground truth. The comparison of world music cultures comprises an exploratory type of research. There is scattered information concerning the ways in which music cultures might be similar, but there is no single source defining all possible relations between them. For computational approaches, it is often necessary to have a ground truth which is used to train and also assess the performance of the algorithms. The notion of music similarity is subjective and considering especially the diversity in world music, creating a ground truth of music similarity judgements is very difficult. Not only is the music diverse and the corpus large, but also music perception varies between listeners with different cultural backgrounds (Stevens, 2012).

Non-robust computational music processing. The automatic extraction of musical attributes is necessary for the large-scale computational analysis of world music. Several computational tools for the analysis of music signals have been designed for the primary aim of Western music analysis (Futrelle and Downie, 2002). This means that the tools may sometimes not be reliable for automatic processing of world music recordings and further developments should be considered. What is more, the extraction of music information from the audio signal can be largely affected by the audio recording quality (Urbano et al., 2014). This is especially a challenge in world music recordings where recording conditions vary vastly and material is preserved with different degrees of fidelity. The majority of world music recordings originate from fieldwork, where continuous audio streams need to be further segmented and curated (either manually or automatically). The evaluation of audio descriptors becomes an essential task in large-scale computational analysis (Panteli and Dixon, 2016).

Limitations of computational music content description. Music descriptors extracted automatically from the audio signal are unable to model the same properties as music descriptors extracted manually by human experts. Computational approaches can more accurately capture low-level characteris-

tics of the audio signal whereas manual approaches can more reliably describe high-level features such as aspects of music perception and cognition. For example, an instrument classification system built for manual annotation referred to instrument properties like 'directly struck' and 'indirectly struck' idiophone (von Hornbostel and Sachs, 1961). In automatic instrument classification, algorithms are trained on features capturing low-level characteristics of the signal for example the 'zero-crossing rate' and 'Mel frequency cepstrum coefficients' (Aucouturier et al., 2005) and higher level classification, such as by instrument type, is performed by learning mappings from the low-level to high-level features. The limitations of computational music description in capturing high-level music properties should be taken into account.

Missing context. The analysis of audio recordings from large music archives has great potential via the application of music information retrieval and data mining technologies. However, information extracted solely from the audio signal is incapable of capturing all the aspects of the practice of a music tradition. Music context often lies beyond the audio signal and understanding this context requires processing other forms of music representation not captured by the algorithms and tools reviewed in this study. The computational study of world music can benefit from the incorporation of additional musical context, for example, music notation, social context, and experts' knowledge and analyses. For example, introducing a music ontology framework (Raimond et al., 2007) covering aspects of world music could contribute significantly to the missing context of audio recordings.

Cultural bias. A cultural bias could affect many aspects of a particular study, from the point of acquiring and selecting data, which features to extract or annotate, which (mathematical, behavioural, computational, cognitive) model to use, and how to interpret the results. The risk of cultural bias is particularly high for the study of world music which includes many different cultures a scholar might not be familiar with.

2.6 Discussion

In the sections above comparative studies with manual and computational approaches were reviewed and criticism and challenges involved in a large-scale computational study were discussed. Below I summarise the conclusions and directions for future work, some of which are adopted for the methods employed in this thesis.

Computational approaches to music corpus analysis have mainly focused on Western popular music (Serrà et al., 2012; Shalit et al., 2013; Mauch et al., 2015b). Computational approaches that have considered world music have ei-

ther used a relatively small and geographically restricted corpus (e.g., less than 1000 recordings to compare African scales (Moelants et al., 2009) or aimed to answer different research questions (e.g., which audio features are most suitable for world music classification (Gómez et al., 2009; Kruspe et al., 2011)). Manual approaches that focus on world music are usually restricted to relatively small datasets (with the exception of Lomax (1976) and Savage (2017) analysing more than 4000 recordings the remaining approaches have studied corpora of less than 1000 recordings (Rzeszutek et al., 2012; Brown et al., 2014; Savage and Brown, 2014; Le Bomin et al., 2016).

The largest corpora in comparative music research have been considered in studies analysing music notation (e.g., almost a million scores were analysed to study the distribution of pitch intervals in classical music (Zivic et al., 2013)). However, while music structure is well represented in music notation, acoustic and performance-style characteristics are not captured. What is more, music notation does not exist in all world music cultures and different notation languages and formats across different styles make the comparison difficult. Therefore a world music comparison based on audio recordings is more plausible in this case.

Given the corpora and methods used in both manual and computational approaches to music corpus analysis as shown in Tables 2.1 and 2.2, and the corresponding criticism as explained in Section 2.4, the following issues need to be addressed for future computational studies.

The majority of the criticism for both manual and computational approaches has focused on the sample not being representative for the research question under investigation (see (Fink, 2013) for the review of (Serrà et al., 2012)), the sample size not being large enough for statistical significance of the findings (see (Nettl, 1970) for the review of (Lomax, 1976)), and the sample not being inclusive of all music cultures of the world (see (Clarke, 2014) for a review of (Savage and Brown, 2014)). As discussed in Section 2.5, the sample size can be maximised by combining recordings from distributed sources and collections and sampling methods can be employed to balance the corpus. The selection criteria to ensure the collection is representative with respect to music style are fulfilled if additional metadata are available, for example, the geographical origins, the language and culture of the performers, the year it was recorded or the era of the music it represents, as well as the primary purpose of the fieldwork study or recording collection.

Criticism of computational approaches raised the issue of the automatically extracted features not being suitable to capture meaningful music attributes (see (Fink, 2013) for a review of (Serrà et al., 2012) and (Underwood, 2015) for a review of (Mauch et al., 2015b)). What is more, for both manual and computational approaches the set of music descriptors has been criticised for

not being complete, i.e., not capturing all essential information about the music in comparison (see (Underwood, 2015) for a review of (Mauch et al., 2015b), (Nettl, 1970) for a review of (Lomax, 1976)). The audio features need to be perceptually evaluated or otherwise demonstrated to be meaningful and a thorough list of necessary music descriptors should be developed. An alternative solution could be to not rely solely on a set of features, e.g. derived from the music notation (where performance-specific characteristics are missing), or audio signal (where high-level or perceptual features are difficult to capture), but to combine both notation, audio, and metadata information for a more balanced study of world music. For example, semi-automatic approaches where manual annotations complement automatically extracted features (Cabrera et al., 2008; van Kranenburg et al., 2010) could provide a better representation of the music that could also partly scale to larger corpora. In addition, approaches that learn from weakly labelled data (e.g. using metadata as weak labels) could also be used to extract more reliable high-level MIR features.

Large-scale music comparisons and evolutionary analyses require advanced computational methods. Extra care needs to be taken to not violate assumptions of the underlying statistical tests (see (Underwood et al., 2016) for a review of (Mauch et al., 2015b)). What is more, a good understanding of the musical characteristics of the corpus is required by the person conducting the research to avoid biasing the methodology or the interpretation of the results (see for example the Western bias remark by Fink (2013) in Section 2.4). Conclusions are more likely to be reliable if validated by experts in other disciplines including musicology, biology, statistics, history, and anthropology.

2.7 Outlook

The fields of musicology and MIR have set the grounds for large-scale music corpus studies. By reviewing manual and computational approaches I highlighted the advantages and strengths of state-of-the-art studies. Manual approaches benefit from direct expert knowledge but are limited by the time-consuming task of manual annotation. Computational approaches benefit from the efficient automatic music processing but can be limited by the knowledge represented in the derived attributes. Criticism of popular music corpus studies focuses on the suitability and size of the corpus as well as how meaningful and robust the extracted music attributes are.

Taking into account the challenges involved in a large-scale computational analysis of world music and the aforementioned critical remarks I discussed how music corpus studies can be improved in the future. In particular, this thesis contributes to the computational analysis of world music corpora by curating

a large world music corpus and describing different sampling methods (Chapter 3). The robustness of computational tools involved in the automatic analysis of world music recordings is explicitly assessed in Chapter 4 and discussed in relation to the musicological findings presented in Chapter 6. Other computational approaches in this thesis limit the Western music bias by learning directly from low-level representations of the world music recordings (Chapter 7).

Chapter 3

Music corpus

This chapter describes the world music corpus considered throughout this thesis and the sampling methods to derive appropriate corpora for the study of music similarity. Descriptive statistics of geographical and cultural aspects of the corpus such as the country, year, and language distribution of recordings are provided as well as information on their duration and recording quality. Several corpora are derived from this world music corpus to test musicological hypotheses and evaluate MIR algorithms. For the former, a music similarity ground truth is assumed to be represented in some form in the metadata. For example, the model of music dissimilarity developed in Chapter 6 considers the country of the recordings as a proxy to music style and recordings from the same country are expected to have similar music characteristics.

Corpora for testing computational algorithms are derived to meet certain statistical or other criteria, for example, the criterion for having each country represented by a balanced number of recordings (e.g., the dataset used in Chapter 6), or the criterion for having vocal parts in the recording (e.g., the dataset used for training a vocal classifier in Chapter 5). Lastly, additional datasets were created from synthetic audio data for control experiments with data augmentation. These datasets were used to assess the robustness of rhythmic and melodic descriptors to transformations of the audio with respect to world music processing challenges (e.g., the evaluation strategy proposed in Chapter 4).

3.1 Creating a world music corpus

This thesis investigates music similarity in a corpus of recorded material from folk and traditional music styles from around the world. In particular, field recordings collected by ethnomusicologists since the end of the $19th$ century are considered. The music corpus is drawn from two large archives, the Smithso-

nian Folkways Recordings[1] and the World & Traditional music collection from the British Library Sound Archive[2]. Both archives include thousands of music recordings collected over decades of ethnomusicological research.

The initial set of recordings included a total of 71833 recordings, 42651 from a subset of the Smithsonian Folkways Recordings collection and 29182 from a subset of the World & Traditional music collection of British Library Sound Archive as curated for the purposes of the Digital Music Lab project (Abdallah et al., 2017). Some recordings from the Smithsonian Folkways Recordings collection included examples of jazz and blues genres as well as recordings of speech such as poetry. These examples were considered irrelevant to the study of world music and hence excluded in subsequent analysis. In particular, the dataset was filtered on the genre of the recordings (whenever the genre information was available) and excluded the examples mentioned above. This gave a total of 29865 recordings from the Smithsonian Folkways Recordings, a total of 28237 recordings from the British Library Sound Archive and a new grand total of 58102 combining both archives. The corpus of 58102 recordings is termed 'The British Library/Smithsonian Folkways World Music Corpus' or simply 'BLSF' throughout the thesis.

The geographical spread of world music recordings in each collection is described as follows. The subset of 29865 recordings from the Smithsonian Folkways Recordings collection includes a large representation from North America (more than 14000 from the United States and around 1200 from Canada). It also includes around 6200 recordings from Eurasia (1300 from the United Kingdom, 700 from Russia, 400 from France), 3800 recordings from South America (Mexico 500, Trinidad and Tobago 400, Peru 400), 2000 from Asia (Indonesia 400, India 400, Philippines 200), 1800 from Africa (South Africa 200, Ghana 200, Kenya 100), and around 400 from Oceania. The subset of 28237 recordings from the World & Traditional music collection of the British Library Sound Archive consists of a large representation (17000) from the United Kingdom. It also includes around 7300 recordings from Africa (mostly from Uganda 3000, Botswana 1000, and South Africa 900), 2300 from Asia (mostly from Nepal 800 and Pakistan 800), and around 600 recordings from Oceania, North and South America. The geographical distribution of recordings in the world music corpus considered in this study is shown in Figure 3.1

Since the medium of analysis is digitised audio, most of the recordings are dated since the 1950*s*, with the exception of some recordings from the British Library collection dated around 1900 which were digitised from wax cylinders. In

[1] `http://sounds.bl.uk/World-and-traditional-music`, accessed 14th November 2017.

[2] `http://folkways.si.edu/folkways-recordings/smithsonian`, accessed 14th November 2017.

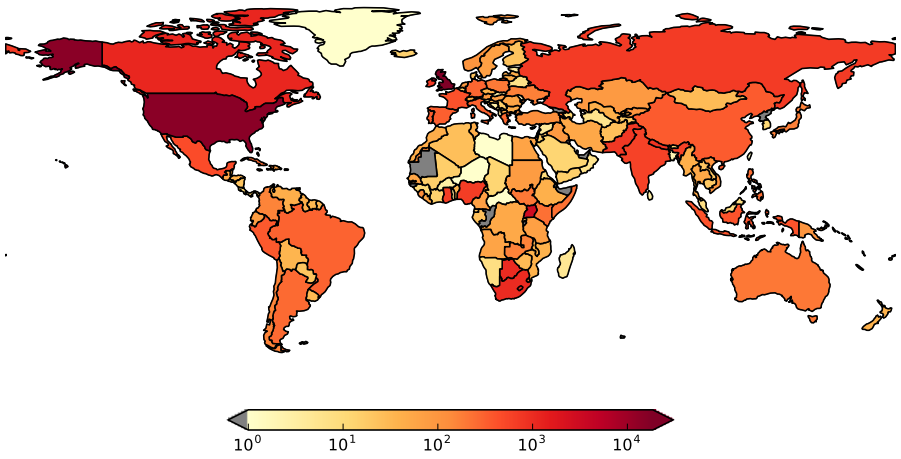


Figure 3.1: The geographical distribution of recordings in the BLSF corpus.

particular, recording dates from the subset of Smithsonian Folkways Recordings span from 1938 to 2014 and from the subset of British Library Sound Archive span from 1898 to 2014. The median year of recording dates in the world music corpus from available information of 54445 examples is 1978 with standard deviation 16 years. The distribution of recording dates by year is shown in Figure 3.2.

Language information is partly provided in the available metadata. In particular, out of 58102 recordings only 33659 include language information, of which only a part (92%) was able to be matched to an official language database (see the curation of language metadata described in Section 3.1.1). From these examples, a total of 103 languages is represented by more than 10 recordings with a mean of 297 and standard deviation of 1964 recordings per language. The most frequent languages in the corpus are English (20050 recordings), Irish (1322), Zulu (778) and Nepali (707). The distribution of the most frequent languages in the corpus (from the available metadata) is summarised in Figure 3.3. The language labels usually denote the language of sung lyrics but sometimes they are also used to denote the language of the performers. No further information was provided in the available metadata to distinguish between the two cases.

The duration of audio recordings from the Smithsonian Folkways Recordings collection is restricted to 30 seconds since only the publicly available 30-second audio previews are used. For the British Library Sound Archive data complete recordings are used which often represent compilations of songs or alternate between segments of speech and music. The duration of recordings from the subset of Smithsonian Folkways Recordings has a mean of 29.0 and standard deviation of 0.8 seconds and from the subset of the British Library Sound Archive

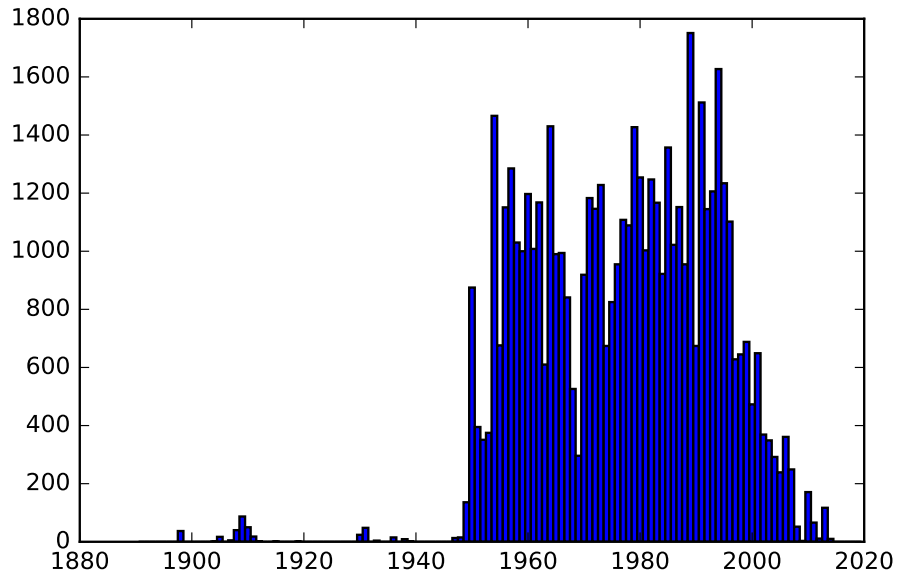


Figure 3.2: The distribution of recording dates by year for the BLSF corpus.

has a mean of 220.0 and standard deviation of 372.6 seconds. In subsequent analysis, for example the developments in Chapter 6, the information extracted from British Library recordings is further processed to only consider short music segments for consistency with the short audio excerpts from the Smithsonian Folkways collection. For the extraction of Mel spectrograms from the 58102 sound recordings, only 25 seconds are considered taken from the middle of each track (Section 4.1.1).

The metadata associated with each music recording include the country where the recording was made and the year it was recorded, the language and sometimes cultural background of the performers, the subject of the music or short description of its purpose, the title, album (if any), and information of the collector or collection it was accessed from. Country information is often more consistent than other culture-related metadata. Several post-processing steps were taken into account to improve the available metadata. These are listed in Section 3.1.1 below.

The BLSF corpus of 58102 recordings as described above with associated metadata is made publicly available for the continuation of this line of research. The audio cannot be shared due to copyrights but the derived (non-reversible) audio Mel spectrograms (Section 4.1.1) are provided instead. The corpus can be accessed at `http://c4dm.eecs.qmul.ac.uk/worldmusicoutliers/BLSF_corpus/`.

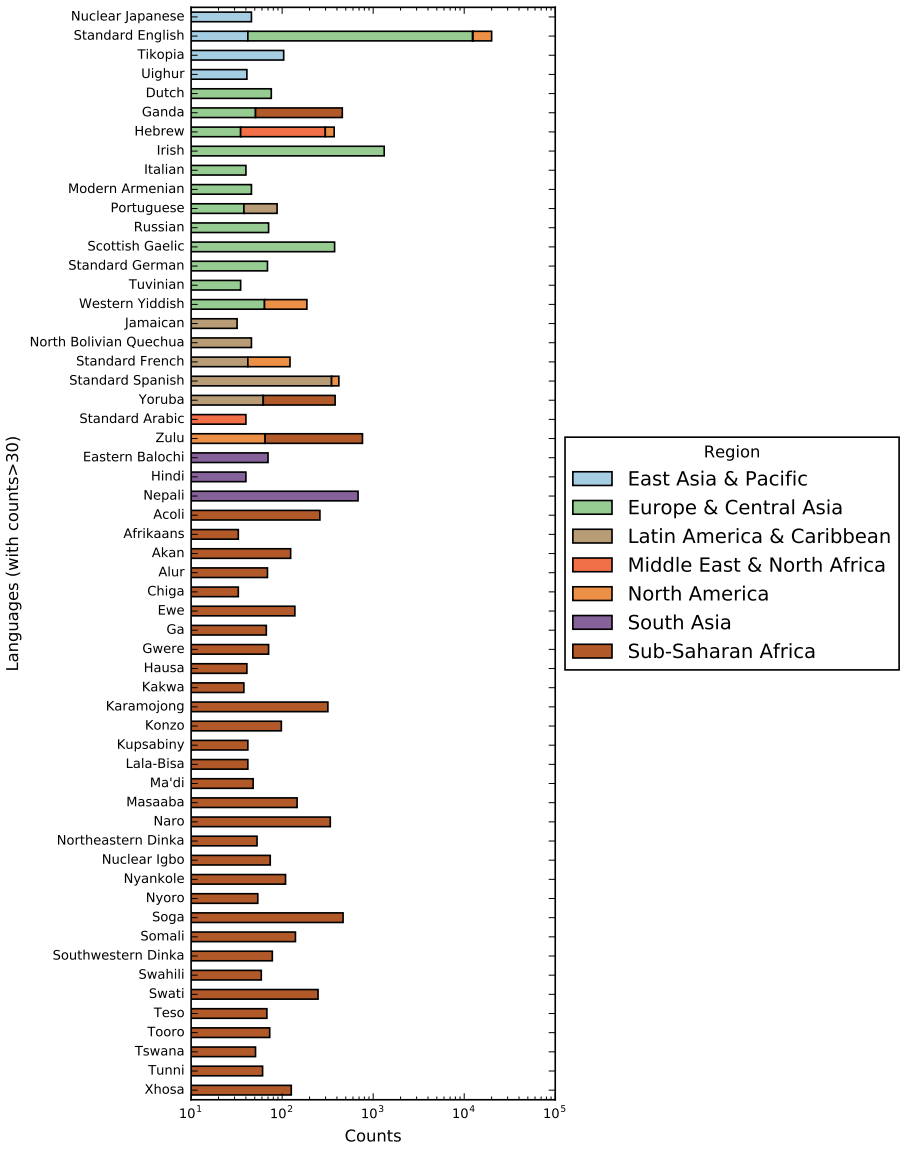


Figure 3.3: The distribution of languages for the BLSF corpus (displaying only languages that occur in more than 30 recordings).

3.1.1 Metadata curation

There are rich metadata associated with each recording in the corpus but the information is not always consistent and there are significant differences between the annotation schemas of the individual music collections. This section describes the process for curating the metadata from the two archives, Smithsonian Folkways Recordings and British Library Sound Archive, to be able to combine them in a unified corpus for automatic analysis.

The metadata taxonomy differs between recordings from the Smithsonian Folkways and British Library Sound archives. Smithsonian Folkways Recordings metadata include the music genre of a recording, typically the social function of the piece, as a separate category. British Library Sound metadata do not include genre information explicitly but sometimes the genre can be derived from the given notes of the collector. In addition, metadata information is registered in different formats, for example Smithsonian Folkways recordings include the year a recording was made as a separate category in the metadata whereas British Library recordings include the full date (day, month, year) the recording was made. Inconsistencies in the metadata are particularly obvious considering the different spelling and labelling in the description of country origins, language and culture of the music, used by different ethnomusicologists and collectors over the years.

Country information is a part of the metadata that is almost always present. Smithsonian Folkways Recordings metadata include country information as a separate category. British Library Sound metadata include country information as part of the location description. This description could be as short as the name of the country, for example 'India', or a full sentence describing the exact location, for example 'In traditional Bhunga mud round house in Dhordo village, Banni area, Kachchh, Gujarat'. The first step in the curation of the metadata is to derive the country information and eliminate possible inconsistencies in the country spellings. This was achieved via a semi-automatic approach.

First, the country information was automatically mapped to the spatial database provided by the R package 'rworldmap: Mapping Global Data, version 1.3.6'[3] to identify entries that are already in a consistent form. Inconsistent entries were inspected manually and adjustments were made to correct the spellings. For example, the spelling 'DR Congo' was corrected to 'Democratic Republic of the Congo' according to the corresponding country entry in the spatial database. Country labels with outdated information such as 'Yugoslavia' or 'USSR' were mapped to updated country labels by manually consulting additional metadata such as the language or culture information of a recording

[3] `http://CRAN.R-project.org/package=rworldmap`, accessed 15th November 2017.

(whenever available). For example, recordings with metadata information 'Yugoslavia' as country and 'Croatian' as language were updated to the country information 'Croatia'. This is likely to be a correct guess most of the time but with no guarantee.

Long sentences describing location in the metadata of British Library Sound Archive recordings were automatically parsed with the 'geocode' function with Google Maps source[4] from the R package 'ggmap, version 2.6.1'[5]. Using this function, the latitude and longitude coordinates of the location were extracted and mapped to a country as identified by the spatial database in the 'rworldmap' package. Again, manual inspection of the data corrected country information from erroneous geocode estimates. Once the country information was corrected and mapped to the spatial database, additional information such as the International Standardisation Organisation (ISO) 3 letter country codes (ISO3), continent from a 6 continent classification system (continent), and region (REGION) (South, 2011) were derived and appended to the metadata.

Information from the recording date of (mainly) recordings from the British Library Sound Archive was processed to derive the corresponding year and decade. The recording dates were stored in different formats, for example, '1965/10/01', '10-03-1978', '1980s', 'ca. 1987'. The year of the recording was automatically extracted from these formats using pattern identification with regular expression techniques[6]. Manual inspection corrected the remaining cases of inconsistent formats such as recording dates labelled as 'unidentified' and 'unknown' that were replaced with 'NA' (denoting 'information Not Available'). From the corrected year information, decade was derived with the simple operation $decade = 10 * floor(year/10)$, for example, the $year = 1988$ would be assigned to the $decade = 1980$.

The language information in the metadata included inconsistent spelling and classification. For example, language labels ranged between describing the exact dialect such as 'Acholi Nyakwai', a dialect of Acholi people in Uganda and South Sudan, the language such as 'Acoli' or the overall language family such as 'Luo'[7]. To process language information, the language database of Glottolog (Hammarström et al., 2017) was consulted. This database includes information on language origins and provides a hierarchical language classification system to distinguish between dialects, languages, and language families. Language information in the metadata of the world music corpus was checked against the Glottolog language database and inconsistent cases were inspected manually.

[4] `http://code.google.com/apis/maps/documentation/geocoding/`, accessed 15th November 2017.

[5] `http://CRAN.R-project.org/package=ggmap`, accessed 15th November 2017.

[6] `https://en.wikipedia.org/wiki/Regular_expression`, accessed 15th November 2017.

[7] `http://glottolog.org/resource/languoid/id/acol1236`, accessed 15th November 2017.

With this semi-automatic process language entries with different spellings, such as 'Acholu' and 'Acholi', were merged to the same language 'Acoli'. Additional information denoting the ISO 3 letter language code and top-level language family was derived and appended to the metadata for those language entries matched in the Glottolog database.

Table 3.1 summarises the metadata information provided with each recording. The curated metadata is made available together with the extracted Mel spectrograms for each recording in the BLSF corpus.

3.2 Derived corpora for testing musicological hypotheses

Even though access to large collections of world music recordings is now feasible, the creation of a representative corpus for the study of music similarity is still challenging. An ideal world music corpus would include samples from all inhabited geographical regions and provide information on the spatio-temporal and cultural origins of each music piece. The samples chosen would have to be sufficient to represent the diversity of styles within each music culture and the corpus as a whole should be a balanced collection of music cultures. Given the archives available today, the challenges in corpus creation involve addressing what defines a good sample, how to balance the diverse styles represented in the collection, how to avoid the Western-music bias and how to maximise the size of the corpus. These challenges have also been the main point of criticism for several music comparative studies (Nettl, 1970; Fink, 2013; Clarke, 2014; Trehub, 2015). In the BLSF corpus described in Section 3.1 there is an unbalanced representation of music cultures, with the majority of recordings originating from Western-colonial areas. The effort to derive world music corpora suitable for the study of music similarity from the currently available metadata is described below.

The following corpora have been created for the research questions addressing singing style similarity (Chapter 5), music dissimilarity and outliers (Chapter 6), and music similarity and expectations (Chapter 7). Chapters 6 and 7 both rely on developing a model for music similarity. However the application changes from detecting distinct recordings also referred to as outliers (Chapter 6) to identifying pairs of similar recordings with respect to audio content and metadata (Chapter 7). To make this distinction clearer, the term 'music dissimilarity' is sometimes used to refer to the developments in Chapter 6 and the term 'music similarity' is used for Chapter 7. These are also the terms used below to define the corpora used in each chapter respectively.

Name	Description	Notes
Title	The title of the music track	
Album_title	The title of the album (if the track belonged to a released album)	Only available for recordings from Smithsonian Folkways Recordings
Artist	The artist of the track	Only available for recordings from Smithsonian Folkways Recordings
Album_genre	Music genres such as 'World' or 'Latin' or 'American Indian'	Only available for recordings from Smithsonian Folkways Recordings
Instrument	The instrument(s) present in the recording	Only available for recordings from Smithsonian Folkways Recordings
Collection	The ethnomusicologist that produced or label company that released the track	
URL	The Universal Resource Locator (URL) of the audio in the original source, Smithsonian Folkways Recordings or British Library Sound Archive	As accessed on 15th November 2017
Year	The year the recording was made or released	
Decade	The decade derived from the year information	
Culture	The culture group of the musicians	Only available for recordings from Smithsonian Folkways Recordings
Orig_language	The original language information before matching it to the Glottolog database	
Language	Language names derived from Glottolog database matching	
Language_iso3	ISO 3 letter language code derived from Glottolog database	
Language_family	The language family derived from Glottolog database	
Region_cultural	Geographical regions from a classification system of 7 broad cultural areas	
Country	The country where the recording was made	
Country_iso3	ISO 3 letter country code derived from rworldmap spatial database	
Region	The region of the recording's location choosing from 6 major geographical areas	
Continent	The continent of the recording's location choosing from 5 continents	Europe and Asia are considered as the same continent 'Eurasia' (South, 2011)
Latitude	Latitude coordinates derived from the recording's location	If only the country name is available it denotes the latitude centroid of the country
Longitude	Longitude coordinates derived from the recording's location	If only the country name is available it denotes the longitude centroid of the country
Orig_location	The original recording location entry before post-processing to extract countries, regions, and continents	
Description	Additional notes on the subject or context of the music	
Archive	The archive that holds the original recording, Smithsonian Folkways Recordings or British Library Sound Archive	
ID_catalog	Identifier of the recording entry in the catalogue of the corresponding archive	
ID_csv	Identifier of the recording in the BLSF corpus	

Table 3.1: The curated metadata available with each recording in the world music corpus.

3.2.1 Corpus for singing style similarity

A dataset of 2808 world music recordings is sampled from the world music corpus for the study of singing style similarity (Chapter 5) in the following way. First, in order to study singing style characteristics recordings that, according to the metadata, contain vocals as part of their instrumentation are selected. From the two available archives, only the Smithsonian Folkways Recordings includes information on the instrumentation of the recording. Thus, the corpus for studying singing styles is derived exclusively from a subset of the Smithsonian Folkways Recordings collection. From this set, vocal recordings are sampled at random from as many countries as possible for geographical diversity and balanced by selecting a minimum of 40 and maximum of 60 recordings per country.

The final set includes recordings from a total of 50 countries with mean 56 and standard deviation 6 recordings per country. Recordings in this dataset span a minimum of 28 different languages and 60 cultures, but a large number of recordings lack language or culture information. In the assessment of singing style similarity, metadata including information on the country, language, and culture of the recording are considered a proxy for similarity.

3.2.2 Corpus for music dissimilarity

A corpus of 8200 recordings from both the Smithsonian Folkways Recordings and British Library Sound Archive collections is used for studying music dissimilarity and outliers (Chapter 6). A corpus is created by sampling recordings based on country information, the most consistent culture-related information in the associated metadata (Section 3.1.1). Similar to the singing style similarity corpus (Section 3.2.1), recordings are sampled from as many countries as possible to ensure a good geographical spread. A minimum requirement of $N_{min} = 10$ recordings from each country and maximum of $N_{max} = 100$ is set. Setting the minimum to 10 recordings is a trade-off between allowing under-represented areas to be included in the dataset and having a sufficient number of samples for each country. Although a sample of 10 recordings is too small to represent the diversity of music styles within a country, raising this minimum to e.g. 50 would exclude many of the countries and would limit the geographical scope of the study. Setting the maximum to 100 recordings prevents the over-represented areas from dominating the corpus. If a country has less than 100 recordings then all recordings (denoted N with $N_{min} < N < N_{max}$) from this country are added to the corpus. If a country has more than 100 recordings then $N = 100$ recordings are sampled at random from this country.

Given the above criteria, the final dataset consists of a total of 8200 record-

ings, 6132 from the Smithsonian Folkways Recordings collection and 2068 from the British Library Sound Archive collection. The recordings originate from 137 countries with mean 59.9 and standard deviation 33.8 recordings per country. A total of 67 languages is represented by a minimum of 10 recordings, with a mean of 33.5 and standard deviation of 33.5 recordings per language. The recordings span the years between $1898 - 2014$ with median year 1974 and standard deviation of 17.9 years.

3.2.3 Corpus for music similarity

A large dataset is sampled from the world music corpus to train an automatic music tagging system with deep neural networks for world music similarity. This dataset combines recordings from the Smithsonian Folkways Recordings and the British Library Sound Archive collections. For the task of learning music similarity from a variety of tags, metadata providing information on the country, language and decade of the recording are considered. To avoid having many tracks from the same country dominating the corpus a maximum of 1000 recordings per country, randomly selected, is kept. From the remaining dataset, recordings are chosen to maximise the occurrence of each tag in the training set. In particular, the tags considered in this experiment need to occur in a minimum of 50 recordings from the corpus.

The final dataset consists of a total of 18054 recordings. There are 80 countries, 33 languages, and 11 decades represented in this dataset. There is a mean of 210 and standard deviation of 208 recordings per country with large representations (more than 500 recordings) of the countries Botswana, Canada, South Africa, Uganda, United States of America, and United Kingdom. Each language tag occurs in a mean of 160 and standard deviation of 203 recordings. The most frequent languages in this dataset are English (1059 recordings), Nepali (533), Zulu (498), and Irish (409). The recordings span the decades between $1900 - 2010$ with small representation for the decades before 1950 and after 2000.

3.3 Derived datasets for testing computational algorithms

The following datasets have been created for the evaluation of MIR algorithms.

- A total of 2170 recordings from the Smithsonian Folkways Recordings collection was sampled for assessing different learned feature spaces as part of the development of a music dissimilarity model (Chapter 6) as

described in Panteli et al. (2016a). In particular, 70 recordings were sampled at random from each of 31 countries from around the world. In this case, country information is considered a proxy for music similarity.

- For the purpose of training a vocal/non-vocal pitch contour classifier as explained in Chapter 5, pitch contours are annotated for a set of 30 world music recordings. Pitch contour annotations are created using the Tony software (Mauch et al., 2015a). Only the lead vocal contour is annotated in each recording. The set of recordings consists of 3 samples from each of 10 countries and is a subset of the Smithsonian Folkways Recordings collection.

3.4 Other datasets

The following datasets have been created for the primary goal of evaluating rhythmic and melodic descriptors.

- A dataset of 3000 synthetic melodies is derived by transforming the audio of 30 melodies in 100 different ways. The transformations include changes in pitch, tempo, recording quality, and instrumentation. The original melodies are given in notation form, in either Western scores, or a list of fundamental frequency estimates and their corresponding time stamps as derived from the sound recordings using the Melodia (Salamon and Gómez, 2012) or YIN (de Cheveigné and Kawahara, 2002) algorithms. These melodies are derived from various world music corpora such as Dutch folk and Byzantine music, and capture particularities of world music characteristics such as microtonality and non-Western scales.

- Similar to the synthetic dataset of 3000 melodies, a dataset of 3000 synthetic rhythm examples is created by transforming the audio of 30 unique rhythmic patterns in 100 different ways. The transformations include changes in global and local tempo, recording quality, and instrumentation. The individual rhythms are derived from transcribed rhythmic patterns from various world music corpora such as Latin-Brazilian, African, and North-Indian music. These rhythms capture particularities of world music characteristics including different time signatures such as $\frac{11}{8}$ in North-Indian, $\frac{12}{8}$ in African, and $\frac{6}{8}$ in Latin Brazilian rhythms.

More information on how the above two datasets were created, along with details on the motivation and application, is provided in Section 4.4.2.

3.5 Outlook

In this chapter the creation of the BLSF world music corpus from sound recordings and metadata available from two large archives was presented. The process to curate the metadata and correct inconsistent information was presented and the dataset of both metadata and sound recordings represented by their corresponding Mel spectrograms (Section 4.1.1), is made publicly available. From this world music corpus, several datasets were derived for testing musicological hypotheses and computational algorithms. The sampling methods for the derivation of each dataset and the ways in which music similarity can be inferred from the metadata were explained. The majority of the sampling methods rely on country metadata as country information is the most consistent culture-related information in the corpus. More information on how the derived datasets are used in subsequent analyses can be found in the corresponding chapters (Chapters 4, 5, 6, and 7). Challenges in the creation of a world music corpus were discussed in Section 2.5 and directions for future improvements are presented in Section 8.2.

Chapter 4

Audio features

The developments in this thesis are based on extracting audio features from a sound recording. Audio features in this case denote a post-processed representation of the audio signal with emphasis on revealing relevant musical characteristics. The design of audio features has been a primary focus in MIR but today, data-driven approaches from the fields of machine learning, and especially deep learning, are becoming more popular.

This thesis uses a variety of audio feature extraction techniques: Chapter 5 focuses on the design of features capturing pitch aspects of the singing voice, Chapter 6 combines expert knowledge in feature design with machine learning methods to learn high-level representations, and Chapter 7 follows a data-driven approach to learn audio features from the Mel spectrograms with deep neural networks. In this chapter I review the different approaches to audio feature extraction, discuss their applicability for the analysis of world music corpora, and propose an evaluation strategy to assess the performance of existing audio features. The rhythmic and melodic descriptors found optimal in this evaluation are used for studying music dissimilarity and outliers as described in Chapter 6.

4.1 Time to frequency domain

A digital sound recording can be considered as a time series of pressure deviation amplitudes, that is, a waveform sampled at short discrete moments in time. This is called the *time-domain* representation of the signal, depicting the amplitude of oscillations over time. The *sampling rate*, number of samples per second, determines the rate of amplitude variation that can be captured and bounds the maximum frequency represented in the discrete signal. For example, a sampling rate of 44100 Hz (a typical sampling rate for sound recordings) can only capture oscillations of up to 22050 Hz (the Nyquist frequency) and oscillations faster

than this rate will be represented with lower frequency when sampled.

The time-domain representation of the signal provides a clear view of the amplitude modulations over time, but frequency information, which is important for modelling musical attributes such as the use of pitch over time, is not explicitly modelled. A useful representation for automatic music analysis is therefore the *frequency-domain* of the signal, i.e., the frequency of oscillations over time. These can be modelled via sinusoids of certain magnitude and phase, also called the *spectrum* of the signal, using the Fourier transform. The computation of spectra over successive short audio excerpts is called a *spectrogram*, i.e., a matrix indicating which frequencies are present in the signal at a given time.

The computation of the spectrogram can be derived with the Short-Time Fourier Transform (STFT),

$$X(t,k) = \sum_{n=0}^{T-1} w(n)x(n+tH)\exp^{-2\pi ikn/T} \tag{4.1}$$

where each time-frequency bin $X(t,k)$ represents the magnitude and phase of a sinusoid of relative frequency k/T derived from the time series $x(n+tH)$ of length T starting at position tH, with T, H the frame and hop size as explained below, $w(n)$ the window function, and $k \in (0, 1, ..., T/2)$ with $T/2$ corresponding to the Nyquist frequency. The magnitude information (*magnitude spectrogram* or just *spectrogram* as will be referred to in this thesis) is considered for subsequent music analysis whereas the phase information is often omitted (Müller et al., 2011).

Essential parameters for the computation of a spectrogram are the frame size T (or *window size* as sometimes referred to in this thesis) and hop size H (as used in Equation 4.1). The frame size T defines the number of consecutive time series samples processed to obtain each frequency component. The choice of the frame size is usually a trade-off between frequency and time resolution. Large frame sizes provide good frequency resolution but the precise timing of each frequency component is lost. A short frame size provides poor frequency resolution whereas the timing is well preserved. The hop size H denotes the distance between successive time frames. Usually, the hop size is chosen with slight overlap between successive frames in order to reduce artefacts caused by the windowing function at the frame boundaries. Related research (Esparza et al., 2014) used frame and hop sizes of around 40 ms with 87.5% overlap, respectively, and these are the sizes used in subsequent analysis in this thesis (see also Section 4.4).

4.1.1 Logarithmic frequency representations

Pitch information is known to follow approximately a logarithmic frequency structure (Lindsay and Norman, 1977). Spectrograms for music analysis are often transformed from a linear frequency scale to a logarithmic one reflecting perceptual attributes of pitch. A common transformation, and the one adopted in this thesis, is the Mel scale (Stevens and Volkmann, 1940). The Mel scale is a scale of pitches judged by listeners to be equal in distance from one another. The spectrograms transformed with the Mel scale are referred to as *Mel spectrograms* throughout this thesis.

Several models have been proposed for the estimation of the Mel scale (Stevens and Volkmann, 1940; Cosell and L., 1976; Lindsay and Norman, 1977; Slaney, 1998). The Mel scale denotes a mapping usually linear below 1000 Hz and logarithmic above 1000 Hz using the reference of 1000 Mels at 1000 Hz. In this thesis, the Mel scale as defined in Slaney (1998) and implemented in Librosa software (McFee et al., 2015b) is used. In this definition, linearly-spaced filters of 133.3 Hz bandwidth are used to model frequencies below 1000 Hz and log-spaced filters separated by a factor of 1.0711703 in frequency are used above 1000 Hz. Formally,

$$mel(f) = \begin{cases} \frac{f}{b} & f < 1000 \\ \frac{1000}{b} + \frac{27}{\log(6.4)} \log(\frac{f}{1000}) & f \geq 1000 \end{cases} \tag{4.2}$$

where f denotes the frequency in Hz and $b = \frac{200}{3}$ denotes half the bandwidth of the linear filter. The Mel scale mapping as defined in Equation 4.2 is shown in Figure 4.1. A filterbank matrix of triangular Mel weights is created, and multiplying this matrix with the original linearly-spaced frequency spectrogram results in the weighted average of frequency bands referred to as the Mel spectrogram.

The Mel spectrograms provided with the BLSF corpus (Section 3.1) are computed in the following way. The STFT algorithm (Equation 4.1) is applied to derive magnitude spectrograms with frame size=40 ms and hop size=5 ms. The frequency bins up to 8000 Hz are converted to the Mel scale with Slaney's formula (Slaney, 1998) as described above. The upper frequency bound of 8000 Hz is set to avoid the high frequency noise that is usually present in old recordings. A total of 40 Mel bands are kept.

Another useful transformation is the derivation of a *chroma vector* or *chromagram*. Similar to the Mel spectrogram, the chromagram also represents frequency information on a logarithmic scale over time. However, the mapping from linear to logarithmic in this case is adjusted to the twelve semitones of

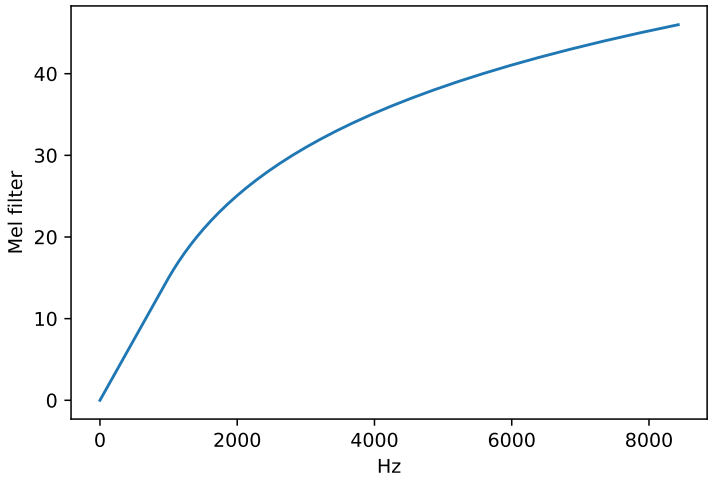


Figure 4.1: The Mel scale mapping for frequencies up to 8000 Hz for the formula defined in Slaney (1998) and implemented in Librosa software (McFee et al., 2015b).

the Western equal-tempered scale. This is usually achieved by applying the constant-Q transform, a Fourier-like transform but with logarithmically spaced filters of constant relative width Q. When first introduced (Fujishima, 1999), chromagrams were derived from the constant-Q transform by the weighted sum of magnitudes over frequency bins belonging to each semitone, and wrapped to a single octave. A more complex definition is that of harmonic pitch class profiles which considers the additional contribution of harmonics (Gómez, 2006). In this thesis a variant of the constant-Q transform is adopted (see also Section 4.4.1).

In this thesis, audio features derived from the chromagrams are used for melodic and harmonic content description whereas audio features derived from Mel spectrograms are (mostly) used for timbral and rhythmic content description (Sections 4.4 and 6.2.3). In Chapter 7, the features extracted from the Mel spectrograms with deep neural networks are used to model general music content description and similarity.

4.1.2 Low, mid, and high level MIR descriptors

A common framework in MIR research distinguishes between low-level, mid-level, and high-level descriptors (Serra et al., 2013). Low-level descriptors capture aspects of the physical attributes of sound and are easily understood by the language of computers and mathematics. High-level descriptors denote abstractions and perceptual music attributes and are the descriptors used by humans. Mid-level descriptors denote an intermediate representation, one that can be

derived by computers and is closer to human music perception.

A desired outcome for automatic music analysis systems is often the derivation of high-level descriptors, i.e., descriptors derived automatically from the audio signal but ones that could be easily understood by a human or used in high-level music description tasks. This is a challenging task for MIR as there is a wide semantic gap between machine and human language (Casey et al., 2008; Wiggins, 2009).

In this thesis I work towards a high-level music description in three ways. a) Custom design of mid and high-level descriptors is considered to approximate perceptual aspects of pitch and the singing voice for the comparison of singing styles in world music (Chapter 5). b) The design of low-level descriptors is considered combined with machine learning approaches to derive high-level representations that can be used to model music dissimilarity (Chapter 6). c) Audio features are learned from low-level representations (in this case Mel spectrograms) using deep neural networks and optimised for the high-level task of predicting relevant music tags for each recording (Chapter 7).

The following sections cover low, mid, and high-level descriptors from the fields of ethnomusicology and MIR for the study of world music.

4.2 Descriptors for world music similarity

Descriptors proposed in the field of comparative musicology for the study of world music cultures include a great deal of expert knowledge to (manually) derive mid and high-level descriptors. One of the first annotation systems is that of 'Cantometrics' (Lomax, 1976). In Cantometrics, descriptors capture aspects of music structure, and performance style, as well as vocal and instrumental elements. In particular, pitch descriptors include properties like melodic range, contour shape, and interval size, rhythm descriptors include properties such as metre and temporal asynchronies in the vocal parts, and phrase attributes such as the length, repetition and instrumentation arrangements. A related study expanding on the concepts of Cantometrics, develops the 'Cantocore' system which places more focus on structural characters rather than the more subjective characters of performance style (Savage et al., 2012). Descriptors in the Cantocore system capture aspects of rhythm, pitch, texture and form and expand the structural descriptors used in Cantomentrics. For example, properties such as the beat sub-division, degree of syncopation, harmonic texture, and type of tonality and mode are added in Cantocore.

For the study of music universals, i.e., the study of musical aspects that are shared across many music cultures of the world, features of sound structure and expressive devices as well as the performance context, content, and behaviour,

are considered (Brown and Jordania, 2011). Qualitative analysis suggests that the use of discrete pitches, octave equivalence, transposition, organisation into phrases, and emotional-intensity factors such as tempo, amplitude, and register modulation are musical attributes shared amongst many music cultures of the world (Brown and Jordania, 2011). A related study expanding on the concept of music universals, uses descriptors from both Cantometrics and Cantocore and finds a set of 18 statistical music universals, i.e., a set of descriptors that is shared amongst many cultures of the world and a network of 10 features that frequently co-occur (Savage et al., 2015a). The 18 descriptors identified as statistical universals include the use of discrete pitches and isochronous beats, and aspects of performance style and social context. The network of 10 descriptors that co-occur link, amongst others, repetitive formal structures, regular rhythms, simple syllabic singing style, use of percussion instruments, group performance and dancing, but exclude pitch attributes.

Studies from the field of MIR that focused on non-Western music repertoires have considered the following audio features. The pitch histogram has been used to study non-Western scales, for example, in Turkish traditional (Bozkurt, 2008), Byzantine (Panteli and Purwins, 2013), Indian traditional (Chordia and Rae, 2007), and African folk (Moelants et al., 2009) music. In each case the pitch resolution of the histogram is tuned respecting the music theory of the tradition. Other features that have been used to compare world music recordings capture low and mid-level aspects of timbre, rhythm and tonality (Gómez et al., 2009; Kruspe et al., 2011; Fourer et al., 2014). For example, timbral features such as MFCCs and spectral centroid, rhythm features such as onset patterns and inter-onset-intervals, and tonal features such as chromagrams and diatonic strength, are used to classify world music recordings by geographical origin.

Audio features considered in this thesis are required to model aspects of similarity in world music. The notion of music style (Section 2.2) is particularly useful in this case: "style can be recognised by characteristic uses of form, texture, harmony, melody, and rhythm" (Sadie et al., 2001). Adopting this definition, descriptors of rhythm, melody, timbre, and harmony are desired. Descriptors of form are ignored in this case since a large part of the recordings from the BLSF corpus consists of only short (30-second) fragments of the total duration of the music tracks (Section 3.1).

Although computational approaches for the study of world music have been considered for over a decade now (Tzanetakis et al., 2007; Gómez et al., 2013), no features have been particularly designed, let alone thoroughly assessed, for the study of world music. In the sections below, methods of evaluation of MIR descriptors are reviewed and an evaluation strategy to assess music descriptors for the study of similarity in world music is proposed.

4.3 On the evaluation of audio features

With the significant number of music information retrieval techniques and large audio collections now available it is possible to explore general trends in musical style evolution (Mauch et al., 2015b; Serrà et al., 2012). Such exploratory studies often have no ground truth to compare to and therefore any conclusions are subject to the validity of the underlying tools. In music content-based systems this often translates to the ability of the audio descriptors to correctly and sufficiently represent the music-specific characteristics. Since this thesis follows also an exploratory approach, evaluation strategies that provide insights on the robustness and relevance of audio descriptors are considered.

The relevance of the features is however not guaranteed, even if a classification task seems successful. For example, unbalanced datasets can lead to high accuracies in genre classification tasks (Sturm, 2013), and high style classification accuracies can be achieved with (only) tempo information (Holzapfel et al., 2011; Dixon et al., 2004) indicating that other music style descriptors used in the classification task had limited relevant contribution. What is more, the music classification system might be picking on aspects irrelevant to the music attributes of the signal. For example, scatter transform features for the classification of recordings by genre were shown to be focusing on frequencies below 20 Hz (i.e., below the common audible range for humans) (Rodriguez Algarra et al., 2016). Timbral features such as MFCCS used for music recommendation were shown to be capturing instrumentation, production and mastering effects, associated with a given album or artist (Flexer and Schnitzer, 2010).

A number of studies have addressed robustness issues of automatic music processing systems in the design of the audio features. For example, rhythmic descriptors have been designed to achieve partial (Holzapfel et al., 2011; Esparza et al., 2014) or complete (Holzapfel and Stylianou, 2011) tempo invariance. In these cases, rhythm information is captured via the onset function, and tempo invariance is achieved by scaling the results with respect to a tempo estimate (Holzapfel et al., 2011; Esparza et al., 2014) or by applying the scale transform that is robust to tempo changes as applied by Holzapfel and Stylianou(2011). Melodic descriptors have been designed for tempo and/or key invariance (Walters et al., 2012; Bertin-Mahieux and Ellis, 2012; Van Balen et al., 2014). Tempo invariance is achieved by using beat-aligned pitch descriptors (Walters et al., 2012; Bertin-Mahieux and Ellis, 2012) and key invariance is achieved by circularly shifting the pitch classes to a reference key (Van Balen et al., 2014).

Other approaches assess robustness via data augmentation techniques. Data augmentation aims at systematically transforming the training examples in a way that encourages a system to learn abstractions that are invariant to these

transformations. It is particularly useful when custom feature design is not desirable nor available. Data augmentation is a popular approach in computer vision with deep learning architectures. In this case, transformations of the image, such as rotation, occlusion, and mirroring, are considered to make the automatic identification system invariant to these transformations. In a similar way, data augmentation techniques in MIR alter the properties of the audio signal to increase the robustness of the automatic music analysis system.

Data augmentation techniques have been considered for the evaluation of MFCCs and chroma features (Urbano et al., 2014). In particular, the robustness of these features was assessed with transformations of the sampling rate, codec, bitrate, frame size and music genre. Using performance statistics in a simulated music genre classification experiment the authors concluded that chroma features are more robust with respect to codec and bitrate than MFCCs (a trivial result considering that chroma features depend on perceptual representations of tonality whereas MFCCs on physical aspects of the signal).

Additional frameworks have been developed for assessing the robustness of audio features with respect to the recording quality. The Audio Degradation Toolbox (ADT) (Mauch and Ewert, 2013) provides a variety of audio effects such as reverb, background noise, pitch and tempo shifts, to modify the audio recording quality. A similar approach is the Musical Data Augmentation (MUDA) framework (McFee et al., 2015a), which includes transformations of pitch shifting, time stretching, background noise, and dynamic range compression. Data augmentation methods for the robustness of singing voice detection consider the addition of random noise, pitch shifting, time stretching, loudness scaling, random frequency filtering, and mixing vocal examples (Schlüter and Grill, 2015).

While the above evaluation methods provide a great variety of approaches and ready-to-use solutions with data augmentation, major challenges involved in world music analysis are not explicitly modelled. For example, pitch shifts in MUDA framework assume intervals relative to the size of a semitone whereas many world music cultures exhibit microtonality (Gómez et al., 2009). In MUDA, background noises of 'subway, crowded concert hall, and night-time city noise' are considered. These effects are unlikely to occur in world music recordings but other types of background noise are usually present, for example, the noise of wind or rain, cars passing by, dogs barking, and crowds applauding or cheering.

In the section below an evaluation strategy based on data augmentation is proposed. The evaluation assesses the robustness of rhythmic and melodic descriptors for the task of world music similarity. Earlier the concept of music style was presented and features capturing aspects of rhythm, melody, timbre, and

harmony were considered relevant for this task. The timbral and harmonic features considered in subsequent analysis (Chapter 6) are derived from MFCCs and chromagrams. These are commonly used features in MIR and their robustness has been assessed in previous studies (Urbano et al., 2014; Walters et al., 2012). The development of rhythmic and melodic descriptors can be more challenging as it involves capturing perceptual attributes over longer time frames (e.g., 8-second frames used for rhythmic descriptors (Esparza et al., 2014) compared to less than 0.05-second frames for timbral descriptors (Pachet and Aucouturier, 2004)). Several aspects of the recording quality of world music and the different particularities in world music styles are considered in the data augmentation techniques described below.

4.4 Evaluating rhythmic and melodic descriptors for world music similarity

In this section the evaluation of audio features that can be used for rhythmic and melodic content description and similarity estimation is described. The proposed evaluation framework aims to simulate challenges in the analysis of recorded world music collections, such as processing noisy recordings or audio samples exhibiting a variety of world music style characteristics. In particular, transformations with respect to timbre, recording quality, tempo and key are defined and the invariance of a set of state-of-the-art rhythmic and melodic descriptors is assessed.

To be perceptually valid, and useful in real-world collections, the representations need to be invariant to subtle changes in tempo, key (or reference pitch), recording quality and timbre. Additionally, to be usable in cross-cultural studies, the features need to be agnostic to properties of particular music cultures. For instance, pitch representations should not depend on the 12-tone equal temperament tuning, and rhythm representations should not depend on specific Western metrical structures such as $\frac{4}{4}$ metre.

A controlled dataset of synthetic audio data which allows to systematically vary timbre, tempo, pitch and audio quality, is created for this purpose. The evaluation strategy assesses the robustness of rhythmic and melodic features in music style classification and retrieval experiments using the synthetic audio dataset. The dataset is made freely available, and the optimal rhythmic and melodic descriptors as shown in the results are used to study world music dissimilarity (Chapter 6).

4.4.1 Features

Details of three descriptors from each category (rhythm and melody), chosen from the literature based on their performance on related classification and retrieval tasks are presented below. My implementations of the features below follow the specifications published in the corresponding research papers but are not necessarily exact replicas because some parameters such as the sampling rate, window, and hop size are fixed across all rhythmic and melodic descriptors for comparison purposes.

Rhythm

State-of-the-art rhythmic descriptors that have been used in similarity tasks including genre and rhythm classification (Esparza et al., 2014; Pampalk et al., 2005; Holzapfel and Stylianou, 2011) are considered. These rhythmic descriptors share the general processing pipeline of two consecutive frequency analyses (Scheirer, 1998). First, a spectrogram representation is calculated, usually with frequencies on the Mel scale. The fluctuations in its 'rows', i.e., the frequency bands, are then analysed for their rhythmic frequency content over larger windows. This basic process has multiple variations, some of which are detailed below.

For comparison purposes the sampling rate is fixed at 44100 Hz for all feature calculations. Likewise, the spectrogram frame size is 40 ms with a hop size of 5 ms. All frequency bins are mapped to the Mel scale. The rhythmic periodicities are calculated on 8-second windows with a hop size of 0.5 seconds. In the second step, the periodicities within each Mel band are computed and averaged across all frames in time.

Onset Patterns (OP). The defining characteristic of Onset Patterns is that the Mel spectrogram is post-processed by computing the first-order difference within each frequency band, subtracting the mean of each frequency band and half-wave rectifying the result. The resulting onset function is then frequency-analysed using the discrete Fourier transform (Pohle et al., 2009; Holzapfel et al., 2011; Esparza et al., 2014). The post-processing step of transforming the resulting linear fluctuation frequencies to $\log_2$-spaced frequencies is omitted because is not consistently considered in the implementation of onset patterns (Holzapfel and Stylianou, 2011). The second frame decomposition results in an $F \times P_O$ matrix with $F = 40$ Mel bands and $P_O = 200$ periodicities linearly spaced up to 20 Hz.

Fluctuation Patterns (FP). Fluctuation patterns differ from onset patterns by using a log-magnitude Mel spectrogram, and by the additional application of psychoacoustic models (e.g. loudness and fluctuation resonance models)

to weight perceptually relevant periodicities (Pampalk et al., 2005). The MIR-Toolbox (Lartillot and Toiviainen, 2007) implementation of fluctuation patterns is used with the frame and hop size parameters specified above. The result is an $F \times P_F$ matrix with $F = 40$ Mel bands and $P_F = 1025$ periodicities of up to 10 Hz.

Scale Transform (ST). The scale transform (Holzapfel and Stylianou, 2011), is a special case of the Mellin transform, a scale-invariant transformation of the signal. The scale invariance property is exploited to provide tempo invariance. When first introduced, the scale transform was applied to the autocorrelation of onset strength envelopes spanning the Mel scale (Holzapfel and Stylianou, 2011). Onset strength envelopes here differ from the onset function implemented in OP by the steps of post-processing the spectrogram. In this implementation the scale transform is applied to the onset patterns (OP) defined above. In particular, the OP descriptor as defined above is post-processed via the application of the Mellin transform.

Melody

Melodic descriptors selected for this study are based on intervals of adjacent pitches or 2-dimensional periodicities of the chromagram. A chromagram representation derived from an NMF-based approximate transcription is used (Maruo et al., 2015).

For comparison purposes the following parameters are fixed in the design of the melodic features: sampling rate at 44100 Hz, variable-Q transform with 3 ms hop size and pitch resolution at 60 bins per octave (to account for microtonality), secondary frame decomposition (where appropriate) using an 8-second window and 0.5-second hop size, and finally averaging the outcomes across all frames in time.

Pitch Bihistogram (PB). The pitch bihistogram (Van Balen et al., 2014) describes how often pairs of pitch classes co-occur within a window d of time. It can be represented as an n-by-n matrix P where n is the number of pitch classes and element p_{ij} denotes the count of co-occurrences of pitch classes i and j. In the current implementation, the pitch content is wrapped to a single octave to form a chromagram with 60 discrete bins and the window length is set to $d = 0.5$ seconds. The feature values are normalised to the range $[0, 1]$. To approximate key invariance the bihistogram is circularly shifted to $p_{i-\hat{i},j-\hat{i}}$ where $p_{\hat{i}\hat{j}}$ denotes the bin of maximum magnitude. This does not strictly represent tonal structure but rather relative prominence of the pitch bigrams.

2D Fourier Transform Magnitudes (FTM). The magnitudes of the 2-dimensional Fourier transform of the chromagram describe periodicities in both

frequency and time axes. This feature renders the chromagram key-invariant, but still carries pitch content information, and has accordingly been used in cover song recognition (Marolt, 2008; Bertin-Mahieux and Ellis, 2012). In the current implementation, chromagrams are computed with 60 bins per octave and no beat-synchronisation. The FTM is applied with the frame decomposition parameters stated above. Only the first 50 frequency bins are selected which correspond to periodicities up to 16 Hz.

Intervalgram (IG). The intervalgram (Walters et al., 2012) is a representation of chroma vectors averaged over different windows in time and cross-correlated with a local reference chroma vector. In the current implementation, only one window is used with size $d = 0.5$ seconds, and cross-correlation is computed on every pair of chroma vectors from successive windows.

The emphasis here is placed on the evaluation framework and a baseline performance of (only) a small set of features is considered. The study could be extended to include more audio descriptors and performance accuracies could be compared in order to choose the best descriptor for a given application.

4.4.2 Dataset

A dataset of synthesised audio, which allows to control transformations under which ideal rhythmic and melodic descriptors should be invariant, is used. Real recordings from world music would have made a much more realistic representation of the audio signal challenges compared to synthesised audio. However, certain transformations such as the ones addressing instrumentation robustness for example by substituting certain instruments, are easier to replicate with synthesised audio signals. In the sections below the dataset of selected rhythms and melodies and a detailed description of their transformations are presented. A summary of this dataset was also presented in Section 3.4.

Material

A set of 30 melodies and 30 rhythms extracted from a variety of musical styles with both monophonic and polyphonic structure is used (Table 4.1). The following melodies are collected: a) MIDI monophonic melodies of classical music used in the MIREX 2013: Discovery of Repeated Themes and Sections task[1], b) MIDI monophonic melodies of Dutch folk music from the Meertens Tune Collections (van Kranenburg et al., 2014), c) fundamental frequency (F0) estimates of monophonic pop melodies from the MedleyDB dataset (Bittner et al., 2014), d) fundamental frequency (F0) estimates of monophonic Byzantine religious music

[1] `http://www.tomcollinsresearch.net/mirex-pattern-discovery-task.html`, accessed 15th November 2017.

(Panteli and Purwins, 2013), e) MIDI polyphonic melodies of classical music from the MIREX 2013 dataset, and f) fundamental frequency (F0) estimates of polyphonic pop music from the MedleyDB dataset. These styles exhibit differences in the melodic pitch range, for example, classical pieces span multiple octaves whereas Dutch folk and Byzantine melodies are usually limited to a single octave range. Pitch from fundamental frequency estimates enables to also take into account vibrato and microtonal intervals. This is essential for microtonal tuning systems such as Byzantine religious music, and for melodies with ornamentation such as recordings of the singing voice in the pop and Dutch folk music collections. Many of the melodies included in this dataset come from Western music styles. This is because at the time of compiling the dataset no other world music collections, to the best of my knowledge, existed with precomputed melody annotations in the form of fundamental frequency estimates or MIDI. In future work, fundamental frequency estimates could be extracted from recordings of world music styles, for example using the MELODIA algorithm (Salamon et al., 2011) and the non-Western music datasets developed in the CompMusic project (`http://compmusic.upf.edu/corpora`).

The following rhythmic sequences are collected which are common in: a) Western classical music traditions (Stober et al., 2014), b) African music traditions (Toussaint, 2003), c) North-Indian and d) Afro-American traditions (Thul and Toussaint, 2008), e) Electronic Dance Music (EDM) (Butler, 2006), and f) Latin-Brazilian traditions[2]. These rhythms span different metres such as $\frac{11}{8}$ in North-Indian, $\frac{12}{8}$ in African, $\frac{4}{4}$ in EDM, and $\frac{6}{8}$ in Latin-Brazilian styles. These rhythmic sequences are represented as binary vectors, with 1 denoting an onset and 0 an offset, and length proportional to the number of beats in a single bar with respect to a given metre. The rhythms for Western, African, North-Indian, Afro-American traditions are constructed from single rhythmic patterns whereas EDM and Latin-Brazilian rhythms are constructed with multiple patterns overlapping in time. The use of a single rhythmic pattern is referred to as 'monophonic' and of multiple patterns as 'polyphonic' for consistency with the melodic dataset.

Transformations

Intuitively, melodies and rhythms retain their character even if the music is transposed to a different tonality, played at a (slightly) different tempo or under different recording conditions. These are variations that are expected to be found in real-world corpora, and to which audio features should be reasonably invariant. As mentioned earlier, key transposition and tempo shift invariance

[2] `http://www.formedia.ca/rhythms/5drumset.html`, accessed 15th November 2017.

Melody		Rhythm	
Description	No.	Description	No.
Dutch Folk (M)	5	Afro-American (M)	5
Classical (M)	5	North-Indian (M)	5
Byzantine (M)	5	African (M)	5
Pop (M)	5	Classical (M)	5
Classical (P)	5	EDM (P)	5
Pop (P)	5	Latin-Brazilian (P)	5

Table 4.1: The dataset of rhythms and melodies transformed for feature robustness evaluation. (M) is monophonic and (P) polyphonic as described in Section 4.4.2.

has been considered for melodic features (Van Balen et al., 2014; Walters et al., 2012; Bertin-Mahieux and Ellis, 2012). Tempo and recording quality invariance has been considered for rhythmic features (Esparza et al., 2014; Holzapfel et al., 2011). In this evaluation strategy, the requirement of invariance to slight changes in timbre is added. The timbre transformations considered in this case are not expected to vastly alter the perception of a rhythm or melody.

Overall, the features are tested for robustness in tempo, pitch, timbre and recording quality by systematically varying these parameters to produce multiple versions of each melody and rhythm (Table 4.2). Only one transformation is applied at a time while the other factors are kept constant. The 'default' version of a rhythm or melody is computed using one of the 25 timbres available, fixing the tempo at 120 bpm, and, for melody, keeping the original key as expressed in the MIDI or F0 values. The dataset is made available online[3].

Timbre (Timb): For a given melodic sequence of MIDI notes or fundamental frequency estimates the audio is synthesised using sine waves with time-varying parameters. The synthesised timbres vary from harmonic to inharmonic sounds and from low to high frequency range. For a given set of rhythm sequences the audio is synthesised using samples of different (mainly percussive) instruments[4]. Beyond the typical drum set sounds (kick, snare, hi-hat), percussive instruments from different music traditions are included such as the Indian mridangam, the Arabic daf, the Turkish darbuka, and the Brazilian pandeiro. Overall, 25 different timbres are modelled for each melody and rhythm in the dataset.

Recording Quality (RecQ): Large music archives usually contain material recorded under a variety of recording conditions, and are preserved to

[3] `http://code.soundsoftware.ac.uk/projects/rhythm-melody-feature-evaluation`
[4] `http://www.freesound.org`, accessed 15th November 2017.

different degrees of fidelity. The Audio Degradation Toolbox (Mauch and Ewert, 2013) is used to create 25 audio degradations that are expected to be found in world music archives. Amongst the degradations, effects of prominent reverb (live recordings), overlaid random noise (old equipment), added random sounds including speech, birds, cars (field recording), strong compression (MP3), wow sampling, and high or low pass filtering (vinyl or low quality microphone) are considered.

Global tempo shifts (GTemp): The tempo changes of up to 20% of the original tempo (in this case centred at 120 bpm) are considered as 'small' variations which leave the perception of melodies and rhythms intact. A total of 25 tempo shifts distributed in the range $[-20, 20]$ (excluding 0) percent slower or faster than the original speed are considered. The tempo shift of 0 (i.e., tempo at 120 bpm) is only considered for the default version of a rhythm or melody.

Key transpositions/Local tempo shifts (KeyT/LTemp): For melodic descriptor robustness the audio is transposed with respect to 25 key transpositions in the range $[-10, 10]$ (excluding 0) semitones from the original key. The key transposition of 0 semitones (i.e., the original key) is only considered in the default version of a melody. These shifts include microtonal intervals, e.g., a transposition of 1.5 semitones up, as one expects to find in world music singing examples. The transpositions are applied to the original melodies prior to chromagram computation. For rhythmic descriptor robustness small step changes of the tempo are considered instead. A *local* tempo change for a duration of 2 (out of 8) seconds centred around the middle of the recording is added. This is common in, for example, performances of amateur musicians where they might unintentionally speed up or slow down the music. Similar to global tempo transformation, 25 shifts in the range $[-20, 20]$ percent slower or faster than the original speed are used.

While the above transformations do not define an exhaustive list of effects and variations found in world music corpora they provide a starting point for assessing feature robustness. For the dataset of 30 rhythms and 30 melodies (Table 4.1) and the above mentioned 4 transformations with 25 values each (Table 4.2) this results in a total of 3000 transformed rhythms and 3000 transformed melodies.

4.4.3 Methodology

The proposed evaluation strategy aims to assess feature robustness with respect to the transformations and transformation values presented above (Section 4.4.2). An additional experiment tests whether the performance of the features relates to particularities of the music style for the styles presented in

Transformations	Values
Timbre	25 distinct timbres (similar frequency range and instrument)
Recording Quality	25 degradations including reverb, compression, wow, speech, noise
Global Tempo	25 values in $[-20, 20]$ percent deviation from original tempo
Key Transposition	25 values in $[-10, 10]$ semitones deviation from original key
Local Tempo	25 values in $[-20, 20]$ percent deviation from original tempo

Table 4.2: Transformations for assessing feature invariance.

Section 4.4.2. Since the dataset consists of monophonic and polyphonic melodies and rhythms, the last experiment tests whether the features are influenced by the monophonic or polyphonic character of the audio signal.

Robustness evaluation is performed on the dataset of 3000 transformed rhythms and 3000 transformed melodies (Section 4.4.2). Considering the variety of MIR tasks and corresponding MIR models, feature performance is assessed in both classification and retrieval experiments as explained below. In these experiments a variety of classifiers and distance metrics are considered, to cover a wide range of audio feature similarity methods.

First the performance of the features to classify different melodies and rhythms is assessed. To do so four classifiers are employed: K-Nearest Neighbours (KNN) with 1 neighbour and Euclidean distance metric, Support Vector Machines (SVM) with a linear kernel, Linear Discriminant Analysis (LDA) with 20 components, and Gaussian Naive Bayes. A 5-fold cross-validation is used for all classification experiments. In each case the prediction target is one of the 30 rhythm or melody 'families'. For each of the 3000 transformed rhythms or melodies the classification accuracy is computed as a binary value, 1 if the rhythm or melody was classified correctly and 0 otherwise.

As reassuring as good classification performance is, it does not imply that a melody or rhythm and its transformations cluster closely in the original feature space. Accordingly, a similarity-based retrieval paradigm is additionally considered. For each of the 30 rhythms or melodies, one of the 25 timbres is chosen as the default version of the rhythm or melody, and used as the query example. The remaining 2999 candidates are ranked based on their distance to the query and the recall rate of its 99 transformations is computed. Each transformed rhythm or melody is assigned a score of 1 if it was retrieved in the top $K = 99$

	Rhythm			Melody		
Metric	ST	OP	FP	PB	IG	FTM
Classification						
KNN	**0.86**	0.71	0.68	**0.88**	0.83	0.86
LDA	**0.82**	0.66	0.59	**0.83**	0.82	0.82
NB	**0.80**	0.62	0.58	**0.84**	0.76	0.81
SVM	**0.87**	0.66	0.59	0.86	0.86	**0.87**
Retrieval						
Euclidean	**0.65**	0.47	0.42	**0.80**	0.56	0.67
Cosine	**0.66**	0.47	0.42	**0.80**	0.55	0.68
Correlation	**0.66**	0.47	0.42	**0.80**	0.54	0.67
Mahalanobis	**0.61**	0.48	0.40	**0.81**	0.60	0.72

Table 4.3: Mean accuracy of the rhythmic and melodic descriptors for the classification and retrieval experiments.

results of its corresponding query and 0 otherwise. Four distance metrics are compared, Euclidean, cosine, correlation, and Mahalanobis distance.

For an overview of the performance of the features the mean accuracy across all recordings for each classification or retrieval experiment and each feature is computed. To better understand why a descriptor is successful or not in the corresponding classification or retrieval task, the performance accuracies with respect to the different transformations, transformation values, music style and monophonic versus polyphonic character are further analysed. To achieve this recordings are grouped by, for example, transformation, and the mean accuracy for each transformation is computed.

4.4.4 Results

The mean performance accuracy of each feature and each classification or retrieval experiment is shown in Table 4.3. Overall, the features with the highest mean classification and retrieval accuracies are the scale transform (ST) for rhythm and the pitch bihistogram (PB) for melody.

Transformation

The mean accuracy per transformation is computed by averaging accuracies of recordings from the same transformation. Results for the rhythmic descriptors are shown in Table 4.4 and for melodic descriptors in Table 4.5. Results are presented for only the best, on average, classifier (KNN) and similarity metric (Mahalanobis) as obtained in Table 4.3. Onset patterns and fluctuation patterns show, on average, lower accuracies for transformations based on global

Metric	Feature	Timb	GTemp	RecQ	LTemp
Classification					
KNN	ST	**0.98**	**0.90**	**0.93**	0.62
KNN	OP	0.97	0.20	0.92	**0.75**
KNN	FP	0.91	0.18	0.92	0.71
Retrieval					
Mahalanobis	ST	**0.95**	**0.36**	**0.91**	**0.25**
Mahalanobis	OP	0.94	0.00	0.88	0.13
Mahalanobis	FP	0.62	0.01	0.87	0.09

Table 4.4: Mean accuracies of the rhythmic descriptors under four transformations (Section 4.4.2).

Metric	Feature	Timb	GTemp	RecQ	KeyT
Classification					
KNN	PB	0.97	**0.99**	**0.78**	0.76
KNN	IG	0.95	**0.99**	0.62	0.77
KNN	FTM	**0.98**	0.96	0.71	**0.79**
Retrieval					
Mahalanobis	PB	**0.94**	**0.98**	**0.78**	0.53
Mahalanobis	IG	0.70	0.91	0.33	0.46
Mahalanobis	FTM	0.87	0.88	0.57	**0.57**

Table 4.5: Mean accuracies of the melodic descriptors under four transformations (Section 4.4.2).

tempo deviations. This is expected as the aforementioned descriptors are not tempo invariant. In the rhythm classification task, the performance of the scale transform is highest for global tempo deviations (KNN accuracy 0.90 and Mahalanobis accuracy 0.36 highest compared to the other descriptors) but it is lowest for local tempo deviations (KNN accuracy 0.62 and Mahalanobis accuracy 0.25 lowest compared to the other descriptors). This is possibly due to the scale transform assumption of a constant periodicity over the 8-second frame, an assumption that is violated when local tempo deviations are introduced. It is also noted that fluctuation patterns show lower performance accuracies for transformations of the timbre compared to the onset patterns and scale transform descriptors.

Transformation value

An additional experiment investigates whether specific transformation values affect the performance of the rhythmic and melodic descriptors. To analyse this

the classification accuracy is averaged across recordings of the same transformation value. There are 25 values for each of 4 transformations so this results in 100 classification accuracies in total. The following observations can be summarised.

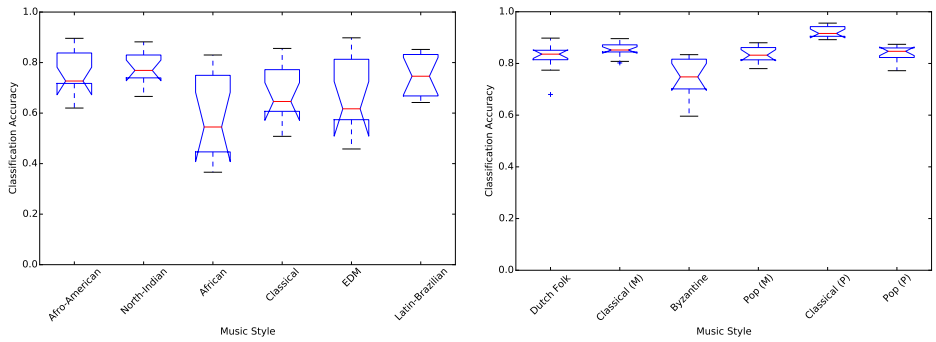
Onset patterns and fluctuation patterns exhibit low classification accuracies for almost all global tempo deviations whereas scale transform only shows a slight performance degradation on global tempo deviations of around $\pm 20\%$. For local tempo deviations, scale transform performs poorly at large local deviations (magnitude $> 15\%$) whereas onset patterns and fluctuation patterns show higher accuracies for these particular parameters. All descriptors seem to be robust to degradations of the recording quality with the exception of a wow effect that causes all rhythmic descriptors to perform poorly. Onset patterns and fluctuation patterns perform poorly also in the degradation of a radio-broadcast compression.

For melody classification, all features perform poorly on key transpositions of more than 6 semitones up and a wow effect degradation. Pitch bihistogram also performs poorly in transpositions between $2.5-5$ semitones down. Intervalgram and Fourier transform magnitudes perform badly for reverb effect degradations and noisy recordings with overlaid wind, applause, or speech sound effects.

Music style

The rhythms and melodies used in this dataset come from different music styles and another question that could be asked is whether the robustness of the features is affected by the music style. To investigate this the classification accuracies are averaged across recordings of the same style. There are 6 styles for rhythm with 500 recordings in each style and likewise for melody (Table 4.1). This gives a total of 6 average classification accuracies for each feature and each classification experiment. Results are summarised in a boxplot as shown in Figure 4.2. Two sets of multiple paired t-tests with Bonferroni correction are also performed, one for rhythmic and one for melodic descriptors, to test whether the classification accuracies per style are significantly different.

The paired t-tests with multiple comparison correction indicated that the majority of pairs of styles have significantly different results at the Bonferroni significance level $alpha = 0.003$ for both the rhythmic and melodic descriptors. In particular the accuracies for classification and retrieval of African rhythms are significantly different from those of all other styles. The classification accuracies for Western classical rhythms are significantly different from the accuracies of all other styles except the EDM rhythms, and the accuracies of North-Indian rhythms are significantly different from those of all other styles except the EDM and Latin-Brazilian rhythms. Low accuracies on average were particularly ob-



(a) Rhythm (b) Melody

Figure 4.2: Box plot of classification accuracies of a) rhythmic and b) melodic descriptors for each style. The accuracies for each style are summarised over all classifiers and all rhythmic and melodic descriptors, respectively.

tained for African and EDM rhythms (Figure 4.2a). For melody, the accuracies for the Byzantine and polyphonic pop styles are significantly different from all other styles. In particular, Byzantine melodies are predicted with, on average, the lowest accuracy. The melodic descriptor that performs particularly badly with respect to Byzantine melodies is the intervalgram and the rhythmic descriptor that performs relatively badly with respect to African and EDM rhythms is the fluctuation patterns. The current results are used as an indication of which styles might possibly affect the performance of the features but the analysis of the intra-style similarity is left for future work.

Monophonic versus polyphonic

The synthesised dataset consists of monophonic and polyphonic melodies and rhythms and another experiment is to test whether the performance of the features is affected by their monophonic or polyphonic character. Similar to the methods described above, classification accuracies are averaged across all monophonic recordings and across all polyphonic recordings. Two paired t-tests are performed, one for rhythmic and one for melodic descriptors, to test whether mean classification accuracies of monophonic recordings are drawn from a distribution with the same mean as the polyphonic recordings distribution. At the $\alpha = 0.05$ significance level the null hypothesis is not rejected for rhythm, $p = 0.25$, but is rejected for melody, $p < 0.001$. The melodic descriptors achieve on average higher classification accuracies for polyphonic ($M = 0.88$, $SD = 0.02$) than monophonic recordings ($M = 0.82$, $SD = 0.04$).

4.4.5 Discussion

This study analysed the performance accuracy of rhythmic and melodic features under different transformations, transformation values, music styles, and monophonic versus polyphonic structure. The scale transform achieved the highest accuracy for rhythm classification and retrieval, and the pitch bihistogram for melody. The scale transform is less invariant to transformations of the local tempo, and the pitch bihistogram to transformations of the key. The rhythmic and melodic descriptors are not invariant to music style characteristics and the performance of particularly melodic descriptors is influenced by the pitch content being monophonic or polyphonic.

This evaluation was performed on a dataset of synthesised audio. While this is ideal for adjusting degradation parameters and performing controlled experiments like the ones presented above, it may not be representative of the analysis of real-world music recordings. The latter involve many challenges, one of which is the mix of different instruments which results in a more complex audio signal. In this case rhythmic or melodic elements may get lost in the polyphonic mixture and further pre-processing of the spectrum is needed to be able to detect and isolate the relevant information.

Results are based on the analysis of success rates on classification and retrieval tasks. This gives an overview of the performances of different audio features across several factors: transformation, transformation value, style, monophonic or polyphonic structure. A more detailed analysis could involve a fixed effects model where the contribution of each factor to the performance accuracy of each feature is tested individually. Alternatively, an analysis of variance (ANOVA) could be used to model the accuracy differences for the various rhythmic and melodic features.

A wide range of standard classifiers and distance metrics was used with default settings. I have not tried to optimise parameters nor use more advanced models since the evaluation had to be as independent of the application as possible. However, depending on the application different models could be trained to be more robust to certain transformations than others and higher performance accuracies could be achieved.

Overall, I have investigated the invariance of audio features for rhythmic and melodic content description of diverse world music styles. A dataset of synthesised audio was designed to test invariance against a broad range of transformations in timbre, recording quality, tempo and pitch. Considering the criteria and analyses in this study the most robust rhythmic descriptor is the scale transform and melodic descriptor the pitch bihistogram. Results indicated that the descriptors are not completely invariant to characteristics of the music style and

lower accuracies were particularly obtained for African and EDM rhythms and Byzantine melodies. The performance of the melodic features was slightly better for polyphonic than monophonic content. The proposed evaluation framework can inform decisions in the feature design process leading to significant improvement in the reliability of the features. The optimal rhythmic and melodic descriptors, scale transform and pitch bihistogram, are used for the study of music dissimilarity in Chapter 6.

4.5 Outlook

In this chapter different methodologies for audio feature extraction were presented. Focus was placed on low and mid-level descriptors from the field of MIR and their applicability to world music analysis. By reviewing approaches from ethnomusicology, where high-level music descriptors are defined and manually annotated, a set of descriptors capturing aspects of rhythm, melody, harmony, and timbre was considered necessary for studying similarity in world music. Techniques for the evaluation of audio features, and more general, automatic music analysis systems were considered.

A new evaluation strategy was proposed to assess robustness of rhythmic and melodic descriptors for the study of world music similarity. This evaluation strategy used data augmentation to create a synthetic dataset of rhythms and melodies simulating some of the challenges found in world music corpora. The evaluation assessed the robustness of three melodic and three rhythmic descriptors and revealed the optimal rhythmic and melodic descriptors for the tasks of music classification and retrieval. These descriptors are used in subsequent analysis of music dissimilarity in a large world music corpus (Chapter 6).

There are numerous methods for audio feature extraction. In the developments of this thesis I consider different approaches, ranging from custom feature design capturing explicit pitch properties (Chapter 5) to complex models automatically learning high-level attributes (Chapter 7). The former requires a relatively good knowledge of the corpus and its essential musical characteristics. The latter requires large amounts of training data and computational power. As more recorded music collections become available, and the advances of feature learning methods achieve accuracies comparable to expert knowledge, custom feature design could be replaced by data-driven approaches. World music however, is still an under-explored topic in the field of MIR, and the consideration of different approaches combining expert knowledge and data-driven learning as mentioned above seems like a suitable path for the current research. Potential directions in future work are discussed in Chapter 8.

Chapter 5

A study on singing style similarity

In this chapter, a study on singing style similarity in world music is considered. The study is performed on a relatively small corpus of music as a preliminary exploration of the world music corpus. The audio features extracted to model singing aspects are designed with expert knowledge consideration and include a range of low and mid-level MIR descriptors. Singing style elements are summarised with dictionary learning methods and similarity is derived via unsupervised clustering. Analyses focus on identifying inter and intra-style singing similarity in world music.

This chapter follows the research published in (Panteli et al., 2017). An additional experiment is conducted to validate the optimal number of dictionary elements. Based on this estimate a new dictionary of contour features is computed and singing style clusters are redefined. Results share similarities to the previous research and expand further on geographical and cultural relations between the derived singing styles.

5.1 Motivation

Singing is one of the most common forms of musical expression. In comparative musicology the use of pitch by the singing voice or other instruments is recognised as a 'music universal', i.e., its concept is shared amongst all music of the world (Brown and Jordania, 2011). Singing has also played an important role in the transmission of oral music traditions, especially in folk and traditional music styles (Bayard, 1950; van Kranenburg et al., 2013). In this chapter a cross-cultural comparison of singing styles is considered using low and mid-level

descriptors to extract pitch information from sound recordings.

In the field of MIR, research has focused on the extraction of audio features for the characterisation of singing styles in Western (Mauch et al., 2014; Salamon et al., 2012), flamenco (Kroher et al., 2014), and Indian Carnatic (Ishwar et al., 2013) music. For example, vibrato features extracted from the audio signal were able to distinguish between singing styles of, amongst others, opera and jazz (Salamon et al., 2012). Pitch class profiles together with timbre and dynamics were amongst the descriptors capturing particularities of a capella flamenco singing (Kroher et al., 2014). Pitch descriptors have also been used to model intonation and intonation drift in unaccompanied singing of a popular Western melody (Mauch et al., 2014) and for melodic motif discovery for the purpose of Indian raga identification in Carnatic music (Ishwar et al., 2013).

Singing style descriptors in the aforementioned MIR approaches are largely based on pre-computed pitch contours. Pitch contour extraction from polyphonic signals has been the topic of several studies (Ozerov et al., 2007; Durrieu et al., 2010; Dressler, 2011; Salamon et al., 2011). The most common approaches are based on melodic source separation (Ozerov et al., 2007; Durrieu et al., 2010) or salience function computation (Dressler, 2011; Salamon et al., 2011), combined with pitch tracking and voicing decisions. The latter two steps are usually based on heuristics often limited to Western music attributes, but data-driven approaches (Bittner et al., 2015) have also been proposed.

In this study the characterisation of singing styles in folk and traditional music from around the world is considered. A set of contour features is developed capturing aspects of pitch structure and melodic embellishments. A classifier is trained to identify pitch contours of the singing voice and separate these from non-vocal contours. Using features describing the vocal contours only a dictionary of singing style descriptors is derived. Unsupervised clustering is used to estimate singing style similarity between recordings and qualitative evaluation refers to culture-specific metadata and listening examples to verify the results.

The contributions of this study include a set of features for pitch contour description, a binary classifier for vocal contour detection, and a dictionary of singing style elements for world music. Studying singing particularities in this comparative manner contributes to understanding the interaction and exchange between world music styles.

5.2 Methodology

The aim is to compare pitch and singing style between recordings in a world music dataset. The methodology is summarised in Figure 5.1. First pitch contours are computed for all sources of a polyphonic signal and pitch descriptors

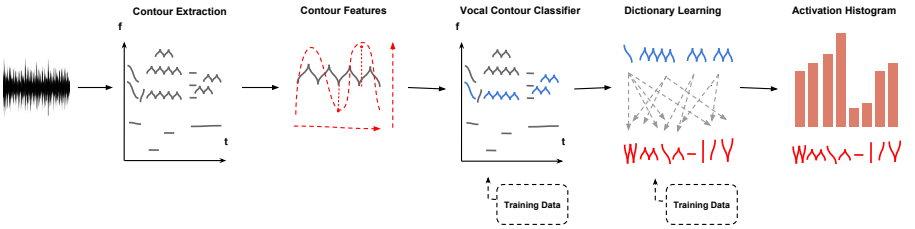


Figure 5.1: Overview of the methodology (Section 5.2): Contours detected in a polyphonic signal, pitch feature extraction, classification of vocal/non-vocal contours and learning a dictionary of vocal features. Vocal contours are mapped to dictionary elements and the recording is summarised by the histogram of activations.

are extracted for each contour (Section 5.2.3). These descriptors are used to train a binary classifier to distinguish between vocal and non-vocal contours (Section 5.2.4). Vocal contours as predicted by the classifier are further processed to create a dictionary of singing style elements. Each contour is mapped to the dictionary matrix and each recording is summarised by the histogram of its contour mappings (Section 5.2.5). Similarity between recordings is modelled via unsupervised clustering and intra- and inter-singing style similarities are explained via references to the metadata and audio examples.

5.2.1 Dataset

The music dataset used in this study consists of 2808 recordings from the Smithsonian Folkways Recordings collection (Section 3.2.1). The publicly available 30-second audio previews and metadata are used and information on the country, language, and culture of the recording is considered as a proxy for similarity. In order to study singing style characteristics recordings that, according to the metadata, contain vocals as part of their instrumentation are selected. The dataset samples recordings from 50 different countries with a minimum of 40 and maximum of 60 recordings per country (mean=56, standard deviation=6). Recordings span a minimum of 28 different languages and 60 cultures, but a large number of recordings lack language or culture information. After extracting vocal contours and filtering out recordings with short or no vocal parts a total of 2766 recordings are kept (see also Section 5.3.1). The top 10 countries, languages and cultures in this dataset can be seen in Table 5.1. The list of recordings and associated metadata in this dataset can be found at `http://github.com/mpanteli/phd_thesis/blob/master/chapter_5/chapter_5_dataset_metadata.csv`.

Additionally, a set of 62 tracks from the MedleyDB dataset (Bittner et al.,

Countries (No.)	Cultures (No.)	Languages (No.)
Argentina (60)	NA (367)	NA (1519)
Puerto Rico (60)	African Caribbean (158)	English (227)
Netherlands (60)	Jewish (130)	Spanish (132)
Nigeria (60)	Anglo-American (88)	Yiddish (64)
Norway (60)	Sami (58)	Dutch (60)
Indonesia (60)	Puerto Rican (55)	Portuguese (58)
India (60)	American Indian (54)	Sotho, Southern (43)
Ghana (60)	African Cuban (51)	Armenian (38)
Germany (60)	Romani (50)	Yoruba (33)
Cuba (60)	Dinka (48)	Japanese (33)

Table 5.1: The top 10 countries, cultures, languages, and their corresponding counts in the dataset of 2766 recordings studied for singing style similarity. 'NA' corresponds to information not available.

2014) containing lead vocals was used as a training set for the vocal contour classifier (Section 5.2.4) and a set of 30 world music tracks containing vocal contours annotated using the Tony software (Mauch et al., 2015a) was used as a test set.

5.2.2 Contour extraction

Pitch contours are extracted with MELODIA (Salamon et al., 2011), a melody extraction algorithm which uses a 'salience function', i.e., a time-frequency representation that emphasises frequencies with harmonic support, and performs a greedy spectral magnitude tracking to form contours. Pitch contours detected in this way correspond to single notes rather than longer melodic phrases. The onset and offset of each contour is estimated based on an energy threshold as found optimal in corresponding evaluation using vocal examples from opera, jazz, pop/rock, and Bossa Nova music (Salamon et al., 2011). The extracted contours covered an average of 71.3% (standard deviation of 24.4) of the annotated vocal contours across the test set (using a frequency tolerance of ± 50 cents). The coverage was computed using the multi-f_0 recall metric (Bay et al., 2009) as implemented in `mir_eval` (Raffel et al., 2014).

5.2.3 Contour features

Each contour is represented as a set of time, pitch and salience estimates. Using this information pitch features are extracted inspired by related MIR, musicology, and time series analysis research. The code is publicly available[1].

[1] `http://github.com/rabitt/icassp-2017-world-music`

Let $c = (\mathbf{t}, \mathbf{p}, \mathbf{s})$ denote a pitch contour for time $\mathbf{t} = (t_1, ..., t_N)$ measured in seconds, pitch $\mathbf{p} = (p_1, ..., p_N)$ measured in Hz, salience $\mathbf{s} = (s_1, ..., s_N)$ measured as the perceived amplitude of frequencies over time, and N the length of the contour in samples. A set of basic descriptors is computed such as the standard deviation, range, and normalised total variation for pitch and salience estimates. Total variation TV summarises the amount of change defined as

$$TV(\mathbf{x}) = \sum_{i=1}^{N-1} |x_{i+1} - x_i|. \tag{5.1}$$

The total variation $TV(\mathbf{p})$ and $TV(\mathbf{s})$ is normalised by $\frac{1}{N-1}$. Temporal information such as the time onset, offset and duration of the contour is also extracted. These descriptors capture the characteristics of the contour at the global level but have little information at the local level such as the turning points of the contour or the use of pitch ornamentation.

The second set of features focuses on local pitch structure modelled via curve fitting. A polynomial y of degree d is fitted to pitch and salience estimates,

$$y[n] = \sum_{i=1}^{d} \alpha_i t_n^i \tag{5.2}$$

for polynomial coefficients α_i and sample $n = 1, ..., N$. The polynomials fitted to the pitch and salience features are denoted $y_p[n]$ and $y_s[n]$, respectively. The coefficients α_{ip} and α_{is} for the polynomials fitted for pitch and salience features respectively, as well as the $L2$-norm of the residuals $r_p[n] = y_p[n] - p_n$ and $r_s[n] = y_s[n] - s_n$ are also stored. The degree of polynomial is set to $d = 5$. These descriptors summarise the local direction of the pitch and salience sequences.

The third set of features models vibrato characteristics. Vibrato is an important feature of the singing voice and the characteristic use of vibrato can distinguish between different singing styles (Salamon et al., 2012). Several methods for vibrato estimation have been proposed in the literature (Herrera and Bonada, 1998; Rossignol et al., 1999). In this thesis vibrato is estimated via a method based on the analytic signal of the Hilbert transform (Hess, 1983) which is a standard approach from the speech recognition literature. Vibrato is modelled from the residual signal between the pitch contour and the fitted polynomial. The residual signal defines fluctuations of the pitch contour not captured via the smoothed fitted polynomial and is thus assumed to carry content of vibrato. Descriptors of vibrato rate, extent, and coverage are extracted from the residual signal as described below.

The residual $r_p[n]$ is approximated by a sinusoid $v[n]$ and amplitude envelope

$A[n]$,

$$r_p[n] \approx A[n] * v[n] = A[n]cos(\bar{\omega}t_n + \bar{\phi}) \tag{5.3}$$

where $\bar{\omega}$ and $\bar{\phi}$ denote the frequency and phase of the best sinusoidal fit. The residual $r_p[n]$ is correlated against ideal complex sinusoidal templates along a fixed grid of frequencies, and $\bar{\omega}$ and $\bar{\phi}$ are the frequency and phase of the template with highest correlation. The amplitude envelope $A[n]$ is derived from the analytic signal of the Hilbert transform of the residual. The sinusoid $v[n]$ is derived by dividing the residual $r_p[n]$ by the amplitude envelope $A[n]$. The frequency $\bar{\omega}$ denotes the rate of vibrato and is constrained to lie within the usual vibrato range of the singing voice and pitch is assumed to fluctuate continuously. Vibrato effects created by the human singing voice are estimated with frequencies around 6 Hz (Prame, 1994). In this implementation the vibrato range is estimated for frequencies ranging between $3 - 30$ Hz. Constraints with respect to fluctuation continuity in time are modelled via the vibrato coverage descriptor C which evaluates the goodness of sinusoidal fit in short consecutive time frames. This is modelled as

$$C = \frac{1}{N}\sum_{i=1}^{N} u_i \tag{5.4}$$

where

$$u_i = \begin{cases} 1, & \text{if } \frac{1}{w}\sum_{k=i-\frac{w}{2}}^{i+\frac{w}{2}-1} |r_p[k] - v[k]| < \tau \\ 0, & \text{otherwise} \end{cases} \tag{5.5}$$

for some amplitude difference threshold τ in this case set empirically to $\tau = 0.25$ (for amplitude values in the range $[-1, 1]$ and maximum difference of 2), time frame of length w centred at sample i, and $r_p[k], v[k]$ the value of the residual and sinusoid, respectively, at sample k. The frame size w is set to the length of half a cycle of the estimated vibrato frequency $\bar{\omega}$. Vibrato effects are considered active in a pitch contour only if the sinusoidal fit is good (according to the τ threshold) for more than 2 full vibrato frequency cycles, i.e., $u_i = 1$ for 4 or more consecutive values of i. If this condition is not met the values of vibrato rate, coverage, and extent are set to 0 for the given pitch contour. The vibrato coverage descriptor is also estimated across three different sections of the contour, namely, the beginning, the middle, and the end, and appended to the feature vector. The process of extracting the vibrato rate and vibrato coverage descriptors from the contour and polynomial input is summarised in Figure 5.2.

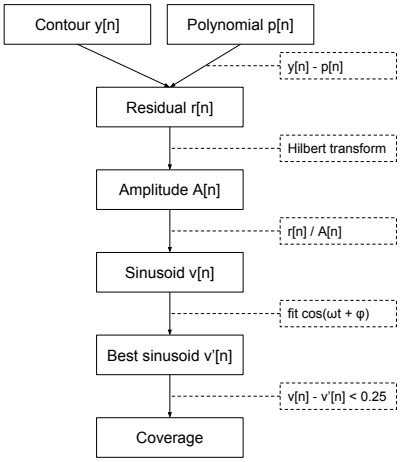


Figure 5.2: The process of deriving the vibrato rate and vibrato coverage descriptors from the residual of the pitch contour and its fitted polynomial using the analytic signal of the Hilbert transform.

Vibrato extent E is derived from the average amplitude of the residual signal,

$$E = \frac{1}{\hat{N}} \sum_{i=1}^{N} u_i A[i] \tag{5.6}$$

for $\hat{N}$ the total number of samples where vibrato was active, i.e., the total number of i such that $u_i = 1$ for u_i as defined in Equation 5.5. The pitch contour $\mathbf{p}$ is reconstructed by the sum of the fitted polynomial, the fitted sinusoidal (vibrato) signal, and some error,

$$p[n] = y_p[n] + E * u[n] * v[n] + \epsilon. \tag{5.7}$$

The reconstruction error ϵ is also included in the set of pitch contour features. An example of the vibrato estimation process from the contour and fitted polynomial is shown in Figure 5.3.

A total of 30 descriptors are extracted summarising pitch content for each contour. Table 5.2 provides an overview of these 30 descriptors. These features are used as input to the vocal contour classifier (Section 5.2.4) and subsequently to learning a dictionary of singing elements (Section 5.2.5).

5.2.4 Vocal contour classifier

A Random Forest classifier is trained to distinguish vocal contours from non-vocal contours using the pitch contour features described above (Section 5.2.3).

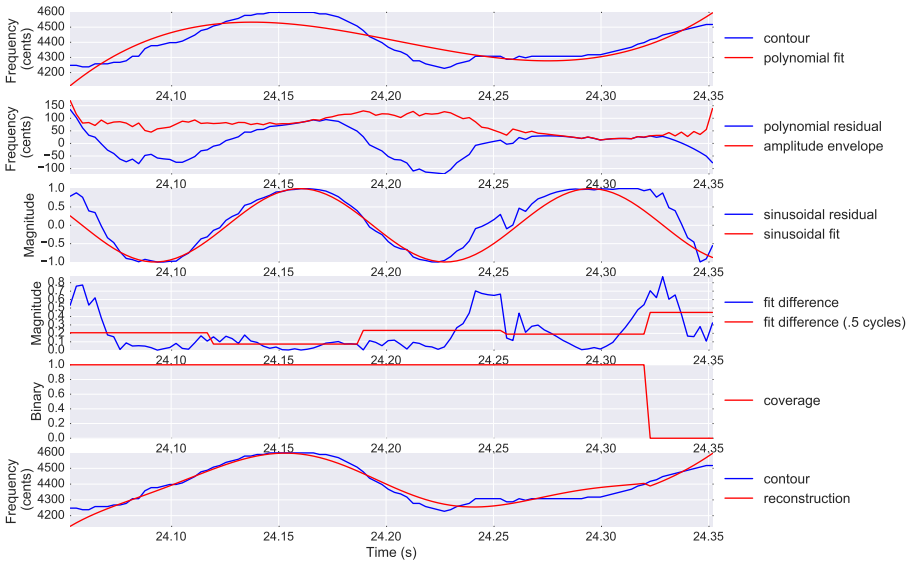


Figure 5.3: Extracting vibrato descriptors from a contour: a) the contour $y_p[n]$ and its fitted polynomial p_n, b) the polynomial residual $r_p[n]$ and the amplitude envelope $A[n]$ derived from the Hilbert transform, c) the sinusoid $v[n]$ and its best sinusoidal fit with frequency $\bar{\omega}$ (the vibrato rate) and phase $\bar{\phi}$, d) the difference between the sinusoid $v[n]$ and best sinusoidal fit per sample and per half cycle windows w, e) the coverage descriptor per sample u_i evaluating the sinusoidal fit difference at the threshold $\tau = 0.25$, f) the original contour and its reconstructed signal.

Index and pitch descriptor	
1. Time onset	16. Pitch polynomial coef. 1
2. Time offset	17. Pitch polynomial coef. 2
3. Duration	18. Pitch polynomial coef. 3
4. Pitch total variation	19. Pitch polynomial coef. 4
5. Pitch standard deviation	20. Pitch polynomial coef. 5
6. Pitch range	21. Pitch polynomial coef. constant
7. Salience total variation	22. Salience polynomial coef. 1
8. Salience standard deviation	23. Salience polynomial coef. 2
9. Salience range	24. Salience polynomial coef. 3
10. Vibrato rate	25. Salience polynomial coef. 4
11. Vibrato extent	26. Salience polynomial coef. 5
12. Vibrato coverage (total)	27. Salience polynomial coef. constant
13. Vibrato coverage (beginning)	28. Pitch polynomial residual
14. Vibrato coverage (middle)	29. Salience polynomial residual
15. Vibrato coverage (end)	30. Pitch contour reconstruction error

Table 5.2: A summary of the 30 descriptors capturing global and local structure of each pitch contour for the study of singing styles in world music.

Training labels are created by computing the percentage a given contour overlaps with the annotated vocal pitch, and labelling contours with more than 50% overlap as 'vocal' (for more details, see Bittner et al. (2015)). The classifier is trained on 62 tracks from the MedleyDB dataset (Bittner et al., 2014) containing lead vocals. The resulting training set contains a total of ≈ 60000 extracted contours, ≈ 7400 of which are labelled as vocal. Hyper-parameters of the classifier are set using a randomised search (Bergstra and Bengio, 2012), and training weights are adjusted to be inversely proportional to the class frequency to account for the unbalanced training set. The classifier is tested on a world music set of 30 tracks (Section 5.2.1). The class-weighted accuracy (Powers, 2011), referred to as F-score hereafter, is used to report the classification results.

5.2.5 Dictionary learning

Given a selection of vocal contours and their associated features a dictionary of the most representative pitch characteristics is computed. Dictionary learning denotes an unsupervised feature learning process which iteratively estimates a set of basis functions (the dictionary elements) and defines a mapping between the input vector and the learned features. In particular, K-means is a common learning approach in image and music feature extraction (Coates and Ng, 2012; Nam et al., 2012).

A dictionary of contour features is computed using spherical K-means, a variant of K-means found to perform better in related research (Dieleman and Schrauwen, 2013). As a preprocessing step, the data is standardised, and whitened with Principal Component Analysis (PCA). A linear encoding scheme is used to map contour features to cluster centroids, obtained by the dot product of the data point with the dictionary matrix.

The number K of clusters used to denote the length of the dictionary is set according to a classification task. Dieleman and Schrauwen (2013) evaluated values of $K \in \{200, 500, 1000\}$ with a tag prediction task using 25863 tracks from the Magnatagatune dataset and the highest accuracies in a variety of experiments were achieved for $K = 500$ and $K = 1000$. In this study the dataset is smaller and values of K are considered in the set $\{10, 100, 200, 500, 1000\}$. The learned dictionaries are assessed with a classification task predicting the country of the recording. The ground truth is derived from the metadata of each recording and represents a total of 50 countries. A Random Forest classifier is trained for the country prediction task given the different dictionaries and the performance is assessed with the F-score metric. The baseline for predicting 1 of 50 countries at random is around 0.02. The dictionary used in subsequent analysis is computed with the optimal value of K as derived from this evaluation.

5.2.6 Singing style similarity

To characterise the singing style of a recording the dictionary activations of its contours are summed and the result is standardised. This is applied to all recordings in the dataset which results in a total of 2766 histograms with K bins each (for K as found optimal in the dictionary learning process, Section 5.2.5). Using these histograms K-means clustering is applied to model singing style similarity. The silhouette score (Rousseeuw, 1987) is used to decide the number $\hat{K}$ of clusters that gives the best partition. Each cluster is considered a proxy of a singing style in the music dataset.

5.3 Results

5.3.1 Vocal contour classification

The vocal contour classifier was tested on the set of 30 world music tracks (Section 5.2.1). The (class-weighted) accuracy on this set was 0.74 (compared with 0.95 on the training set), with a vocal contour recall of 0.64 and vocal contour precision of 0.52. This difference in performance can be attributed to differing musical styles in the training and test set - the training set contained primarily Western pop and classical vocals, while the test set contained vocal styles from across the world. Future improvements for the vocal contour classifier are discussed in Section 5.4.

The classifier was applied to detect vocal contours in the full dataset of 2808 recordings. For a given contour, the random forest classifier outputs the probability of the contour being vocal as well as the probability of the contour being non-vocal. Contours for which the probability of belonging to the vocal class was above 0.5 were considered vocal contours. The classifier outputs also the relative importance of the features for distinguishing between vocal and non-vocal contours. From this analysis, the pitch mean, standard deviation, and total variation were amongst the most important features for the classification task. This implies that the absolute frequency and range of the contour as well as the rate of change of the pitch values are important for distinguishing between vocal versus possibly speech and instrumental contours. Other features that received high importance for the classification task were the salience features indicating that the prominence of the contour compared to other signals in the audio mix is a distinguishing factor.

False negatives (i.e., vocal contours undetected by the classifier) are of little consequence for subsequent analysis, as long as there are a sufficient number of vocal contours to describe the track. False positives, on the other hand, constituting 48% of the predicted vocal contours, do affect this analysis as dis-

K	F-score
10	0.046
100	0.083
200	0.088
500	**0.091**
1000	0.081
Baseline	0.020

Table 5.3: Classification results for different values of K in the computation of a dictionary of contour features with spherical K-means.

cussed in Section 5.3.3. The vocal contour classification reduced the dataset from 2808 recordings to 2766 recordings containing vocal contours. Out of the 2766 recordings, the maximum number of extracted contours for a single track was 458, and the maximum number of extracted vocal contours was 85. On average, each track had 26 vocal contours ($\pm$14), with an average duration of 0.6 seconds. The longest and shortest extracted vocal contours were 11.8 and 0.1 seconds respectively.

5.3.2 Dictionary learning

Dictionary elements were derived from contour features via spherical K-means. The optimal value of K was assessed via a classification task predicting the country of each recording (Section 5.2.5). The results from this classification are presented in Table 5.3. The highest accuracy was achieved for $K = 500$. The dictionary of pitch contour features used in subsequent analysis is thus derived with $K = 500$ elements.

5.3.3 Intra- and inter-style similarity

Using vocal contour features a dictionary of 500 singing elements was built (Section 5.2.5) and a histogram of dictionary activations was computed for each recording. Similarity was estimated via K-means with $\hat{K} = 9$ according to the silhouette score (Section 5.2.6). Figure 5.4 shows a visualisation of the feature space of the recordings using a 2-dimensional t-distributed Stochastic Neighbour Embedding (TSNE) (van der Maaten and Hinton, 2008) and coloured by the cluster predictions. An interactive visualisation of Figure 5.4 can be found at `http://mpanteli.github.io/phd_thesis/chapter_5/TSNE.html`.

Table 5.4 shows the 5 most frequent countries in each cluster. Referring to the metadata of each recording the following observations can be summarised. The majority of singing clusters represented recordings from neighbouring countries or of similar culture or language. For example, cluster 6 grouped mostly

African countries whereas cluster 7 grouped mostly Caribbean countries. Cluster 3 grouped together mostly central European countries such as Germany, Poland, and Netherlands. It also included some music examples from China exhibiting mostly the Western opera singing style. Cluster 9 had some overlap with cluster 3 but additionally grouped together singing examples from the indigenous Finno-Ugric Sami people inhabiting the Arctic area (recorded in Norway) with the indigenous Ona people inhabiting the Patagonian region (recorded in Argentina). Though geographically distinct, the nomadic culture and survival in cold climates are factors that might have contributed to the development of similar singing styles. Cluster 8 grouped together Eastern Mediterranean cultures such as recordings from Italy, Greece, and Israel. It also included Japanese recordings featuring the shakuhachi wind instrument (many of them non-vocal) and Australian aboriginal singing examples. Cluster 2 grouped Latin American countries such as Guatemala and Brazil with French music examples from the Cajun ethnic group as well as music from the Netherlands with South African singing examples in Afrikaans (descendant from Dutch). Clusters 4, 5, and 1 grouped together geographically distinct music. Listening to music examples from these clusters indicated that the grouping criteria were based on attributes of speech and choir singing rather than obvious geographical or cultural links as explained below.

Listening to music examples revealed that some clusters were distinguished by characteristic uses of vibrato, melisma, and slow versus fast syllabic singing. Vibrato can be defined as small fluctuations in pitch whereas melisma is the method of singing multiple notes to a single syllable. It was observed that cluster 9 consisted of slow syllabic singing examples with extensive use of vibrato. In this cluster examples of opera and throat singing techniques were found as well as male (low-frequency) singing voices. With some overlap to cluster 9, cluster 3 also included slow syllabic singing examples with some use of vibrato. Cluster 2 consisted of medium-fast syllabic singing with some use of melisma whereas cluster 8 consisted of medium-slow syllabic singing with extensive use of melisma. In cluster 8 non-vocal examples of the Japanese shakuhachi instrument were also found played with a technique that created effects similar to the singing voice's melisma. Clusters 6 and 7 consisted of fast syllabic singing often with the accompaniment of polyphonic instruments. Cluster 4 consisted of medium-fast syllabic singing and especially choir singing examples with voices overlapping in frequency range creating sometimes roughness or vibrato effects. Cluster 5 included examples of fast syllabic singing especially from dance songs but a lot of these examples captured also prominent instrumental (non-vocal) parts. Cluster 1, the points of which seemed to be slightly disconnected from the other clusters in Figure 5.4, included spoken language examples such as recitation of

Cluster 1 (267)	Cluster 2 (315)	Cluster 3 (173)
United Kingdom (26)	Guatemala (20)	Poland (16)
Argentina (22)	Brazil (13)	Germany (15)
Ireland (20)	Netherlands (12)	Norway (14)
United States (18)	China (12)	Netherlands (12)
Spain (17)	South Africa (11)	China (11)
Cluster 4 (390)	Cluster 5 (334)	Cluster 6 (478)
Armenia (27)	Brazil (16)	Botswana (28)
Sudan (21)	China (14)	Ghana (27)
Nigeria (17)	Hungary (14)	Cuba (24)
Trinidad Tobago (17)	Mexico (13)	Liberia (22)
Colombia (17)	Colombia (13)	Jamaica (20)
Cluster 7 (319)	Cluster 8 (212)	Cluster 9 (278)
Cuba (18)	Italy (18)	Norway (32)
The Bahamas (17)	Greece (16)	Poland (23)
Trinidad Tobago (13)	Israel (12)	Canada (18)
Mexico (13)	Japan (12)	Lesotho (12)
Zambia (12)	Australia (10)	Japan (9)

Table 5.4: The 5 most frequent countries and the number of corresponding recordings (in parentheses) in each singing style cluster.

poems or sacred text.

5.4 Discussion

In this study pitch contour features were extracted for the characterisation of singing styles in world music. These features were used to train a vocal classifier as well as to learn a dictionary of singing style elements. Similarity in singing styles was estimated via an unsupervised K-means clustering method. Results indicated that singing style clusters often grouped recordings from neighbouring countries or with similar languages and cultures. Clusters were distinguished by singing attributes such as slow/fast syllabic singing and the characteristic use of vibrato and melisma.

Further analysis showed that some recordings contained instrumental (non-vocal) or speech contours. The vocal contour classification task can be improved in future work with more training examples from world music, and enhanced classes to cover cases of speech. Contour features such as pitch range and vibrato could possibly distinguish between speech and music contours but it is more challenging to distinguish between instrumental contours that imitate singing, for example the shakuhachi wind instrument with vocal-like melodic phrases as described in the results above. Additional features capturing timbre properties

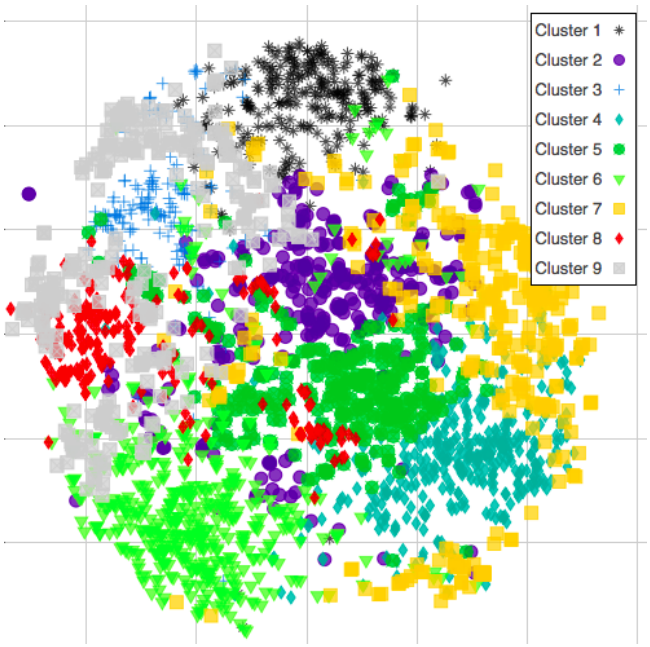


Figure 5.4: A 2D TSNE embedding of the histogram activations of the recordings coloured by the cluster predictions.

can be considered for the vocal contour classifier. While pitch descriptors might have been adequate to model vocal melodic activity, the vocal detection task could be improved with additional features such as MFCCs, spectral centroid, and zero crossing rate, as also proposed in related research (Berenzweig and Ellis, 2001; Murthy and Koolagudi, 2015). Combining timbre with melodic features was also found to be advantageous in distinguishing singing styles of different musical genres (Salamon et al., 2012).

Sub-groups could sometimes be observed within the clusters (e.g., clusters 7 and 9) an indication that the clustering partitions can be further improved. Singing style similarity observations were based on qualitative measures via listening to some examples and visualising the clustered data. Future work could explore further the singing style clusters via a quantitative comparison with the metadata and using feedback from musicology experts. The investigation of singing styles as proposed in this study can provide evidence of interaction and exchange between world music styles.

5.5 Outlook

In this chapter custom feature design was considered to model elements of singing style. The singing clusters revealed interesting findings: attributes such

as slow or fast syllabic singing and the characteristic use of vibrato and melisma distinguish between different singing styles. Clusters often grouped together recordings from neighbouring countries or with similar languages and cultures, for example, a cluster grouped together recordings from mostly Eastern Mediterranean cultures, while another cluster was formed by recordings from mostly European cultures.

With this study the first exploration of similarity relationships in world music was achieved. Amongst the lessons learnt is that training the models with Western music examples might not yield the most accurate results. For example, the vocal contour classifier was trained to distinguish between vocal and non-vocal contours in Western pop melodies but was applied to identify vocal contours in world music recordings. While the classifier achieved an accuracy well above the random baseline, the resulting contours sometimes contained non-vocal parts. This is partly because the instruments found in world music collections cover a wider range of sounds than the ones typically found in Western pop melodies. Training the classifier with particularly music examples from the world music dataset could improve the vocal contour prediction.

Learning the features directly from the world music data could sidestep the Western music bias observed above. This is the approach followed in the next chapters (Chapters 6 and 7) where the methodology is enhanced with feature learning to derive data-driven music descriptors. Chapters 6 and 7 complement the findings of this chapter by expanding the corpus to consider additional music cultures and by broadening the notion of similarity to capture music attributes beyond pitch and singing.

Chapter 6

A study on music dissimilarity and outliers

In Chapter 5 singing style similarity was considered for a small corpus using customly designed audio features. In this chapter, a model of music dissimilarity is developed. The model is applied on a larger and more geographically diverse world music corpus. The audio features derived for this study combine low-level MIR descriptors with machine learning methods to learn higher-level representations directly from the world music data. Music dissimilarity is assessed with an outlier detection technique. Results identify unique musical elements in sound recordings and reveal geographical patterns of music outliers.

This chapter uses the optimal rhythmic and melodic descriptors from the evaluation described in Section 4.4 and the corresponding publication (Panteli and Dixon, 2016). The development of the methodology has been assessed in two separate publications (Panteli et al., 2016a,b) where a feature space for world music similarity was learned from relatively small music corpora. This chapter provides an overview of the above publications, explains some final adaptations to the methodology, and presents the results from the application of the methods to the study of dissimilarity in a large world music corpus.

6.1 Motivation

The history of cultural exchange in the world goes back many years and music, an essential cultural identifier, has travelled beyond country borders. But is this true for all countries? What if a country is geographically isolated or its society resisted external musical influence? Can we find such music examples whose characteristics stand out from other musics in the world? This study focuses on

music dissimilarity and in particular outlier detection.

Outlier detection is a common pre-processing step in the analysis of big data collections (Aggarwal and Yu, 2001). In music, outlier detection can reveal recordings with outstanding musical characteristics, in this thesis referred to as 'music outliers'. The purpose of outlier detection is two-fold. First, it can reveal recordings that need to be filtered out from the collection. For example, a speech or heavily distorted sample could be regarded extreme and undesirable in a collection of sung melodies. Second, given a representative sample of recorded world music, outliers can reveal recordings with outstanding musical characteristics. Tracing the geographic origin of music outliers could help identify areas of the world that have possibly developed a unique musical character.

In previous work I have explored the suitability of audio features for music similarity and content description (Panteli and Dixon, 2016). Audio features for the purpose of studying world music need to be agnostic to style characteristics so that they can generalise to the diversity of music styles. In the evaluation described in Section 4.4, rhythmic and melodic descriptors that are invariant to tempo and pitch transformations and are fairly robust to transformations of the recording quality were found. These features were used in combination with feature learning to assess music similarity in a relatively small world music corpus (Panteli et al., 2016a) as well as to detect and analyse music outliers in a preliminary study (Panteli et al., 2016b). In this study I expand prior work to world music analysis using a larger corpus and evaluating additional audio feature extraction and data mining methods.

Amongst the contributions of this study is the proposed methodology that uses prior knowledge to derive low-level descriptors and combines this with feature learning techniques to adapt the representations to particularities of the world music data. Results from this approach are evaluated quantitatively using metrics to assess classification accuracy and qualitatively via a listening experiment and visualisation of the projected space. This is the first study to investigate outliers in world music on such a large scale. These developments contribute to defining concepts and methods from which future work in the study of large world music corpora can benefit.

6.2 Methodology

The methodology is summarised as follows. For each audio recording in the dataset, music descriptors are extracted by a) filtering out speech segments detected via a speech/music discrimination algorithm, b) extracting low-level audio features capturing aspects of music style, c) applying feature learning to reduce dimensionality and project the recordings into a musically meaningful

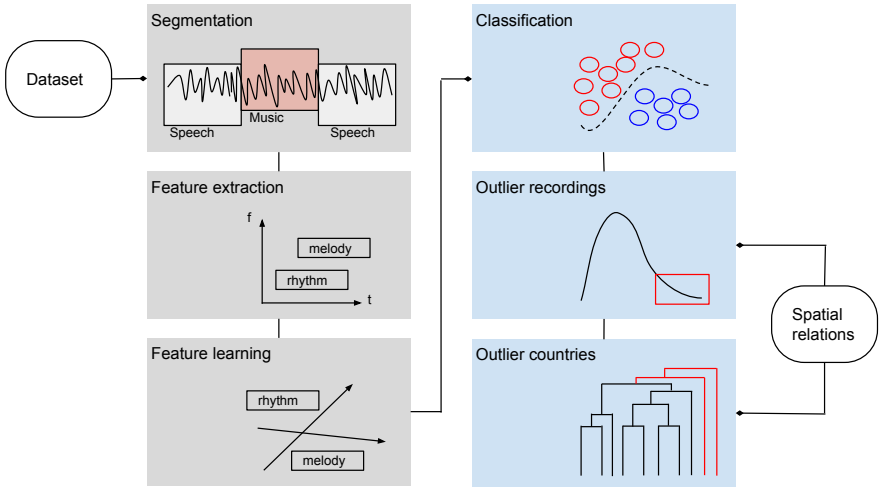


Figure 6.1: Overview of the methodology for the study of music dissimilarity and outliers.

space. A classification task serves to optimise the parameters and assess music similarity in the projected spaces. The optimal feature space is used to identify recordings that stand out with respect to the whole set of recordings (also referred to as outliers). Outliers are detected for different sets of features focusing on rhythm, melody, timbre, or harmony and a combination of these. Spatial information from each country is taken into account to form geographical neighbourhoods and detect 'spatial outliers', i.e., recordings that stand out with respect to their neighbours. A feature representation for each country is extracted by summarising information from its recordings. Hierarchical clustering is used to get an overview of similarity and dissimilarity relationships between countries. The methodology is summarised in Figure 6.1 and explained in detail in the sections below.

In subsequent analysis the country label of a recording is used as a proxy for music style. It is assumed that recordings originating from the same country have common musical characteristics and this is used as ground truth to train the music (dis)similarity model. However, music styles may be shared across many countries and a country may exhibit several music styles. The reason for choosing country as the unit of analysis in this study is two-fold. First, the country label is the most consistent information available in the music metadata compared to, for example, music genre, language, or culture information (see also Section 3.1.1). Second, several studies have considered larger geographical regions (e.g., continents or cultural areas) for the comparison of music styles (Titon et al., 2009; Gómez et al., 2009; Kruspe et al., 2011). Country bound-

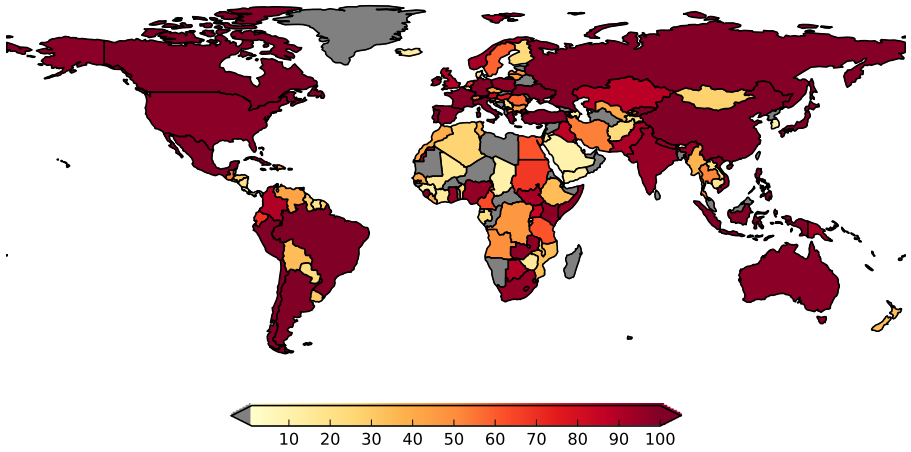


Figure 6.2: The geographical distribution in the dataset of 8200 recordings studied for music dissimilarity.

aries work in a similar way but provide a more fine-grained unit for analysis. Alternative approaches are discussed further in Sections 6.5.2 and 8.2.

6.2.1 Dataset

The dataset of 8200 recordings as described in Section 3.2.2 is used for this analysis. This dataset combines recordings from the Smithsonian Folkways Recordings and British Library Sound Archive and represents a total of 137 countries with a minimum of 10 and maximum of 100 recordings sampled at random from the BLSF corpus. The geographical distribution of this dataset is shown in Figure 6.2. A total of 67 languages is represented by a minimum of 10 recordings, with a mean of 33.5 and a standard deviation also of 33.5 recordings per language. The most frequent languages in this dataset are English (203 recordings), Spanish (147), Swazi (99), Zulu (97), Hassaniya (87), Yoruba (84), and Nepali (81). The recordings span the years between $1898 - 2014$ with median year 1974 and standard deviation of 17.9 years. The list of recordings and associated metadata in this dataset can be found at `http://github.com/mpanteli/phd_thesis/blob/master/chapter_6/chapter_6_dataset_metadata.csv`.

6.2.2 Pre-processing

The dataset consists of field recordings that sometimes mix speech and music segments. Only the music segments are relevant for this study but due to the lack of metadata, speech segments cannot be filtered out a-priori. An essential pre-processing step is therefore the discrimination between speech and music segments. Speech/music discrimination refers to the detection of segment

boundaries and the classification of the segment as either speech or music. This has been the focus of several studies in the literature (Scheirer and Slaney, 1997; El-Maleh et al., 2000; Panagiotakis and Tziritas, 2005) and it was also identified as a challenge in the 2015 Music Information Retrieval Evaluation eXchange (MIREX) (Downie, 2008).

The best performing algorithm (Marolt, 2015) from the MIREX 2015 evaluation is chosen. As part of the MIREX 2015 evaluation, the algorithm was tested on a non-overlapping set of British Library Sound Archive recordings which is very similar to the recording collection used in this study and achieved the highest frame-based F-measure of 0.89. The algorithm is based on summary statistics of low-level features including Mel frequency cepstrum coefficients (MFCCs), spectral entropy, tonality, and 4 Hertz modulation, and it is trained on folk music recordings (Marolt, 2009). This algorithm is applied to detect speech/music segments for all recordings in the dataset and only the music segments are used in subsequent analysis. In case of long audio excerpts the initial music segments up to a total duration of maximum 30 seconds are considered (see the differences in duration between recordings of the Smithsonian Folkways Recordings and British Library Sound Archive collection, Section 3.1).

6.2.3 Audio features

Over the years several toolboxes have been developed for music content description (Tzanetakis and Cook, 2000; Peeters, 2004; Lartillot and Toiviainen, 2007; McFee et al., 2015b). Applications of these toolboxes include tasks of automatic classification and retrieval of mainly Western music (Chapter 2). Audio content analysis of world music recordings has additional challenges. First, the audio material is recorded under a variety of recording conditions (live and field recordings), and is preserved to different degrees of fidelity (old and new recording media and equipment). Second, the music is very diverse and music descriptors designed primarily for Western music might fail to capture particularities of world music styles.

Between specifically designing the music descriptors as in other comparative music studies (Gómez et al., 2009; Kruspe et al., 2011; Zhou et al., 2014) and automatically deriving them from the spectrogram (Hamel and Eck, 2010; Choi et al., 2016) a middle ground is chosen. Expert knowledge is used to derive low-level music representations that are later combined with feature learning methods to adapt the representations to particularities of the world music examples.

State-of-the-art descriptors (and adaptations of them) are selected that aim at capturing relevant rhythmic, timbral, melodic, and harmonic content (Sec-

tion 4.2). In particular, onset patterns with the scale transform (Holzapfel and Stylianou, 2011) are extracted for rhythm, pitch bihistograms (Van Balen et al., 2014) for melody, average chromagrams (Bartsch and Wakefield, 2005) for harmony, and MFCCs (Logan, 2000) for timbre content description. These descriptors are chosen because they define low-level representations of the musical content, i.e., a less detailed representation but one that is more likely to be robust with respect to the diversity of the music styles considered here. In particular, the rhythmic and melodic descriptors selected for this study were chosen based on the evaluation in Section 4.4.

The audio features used in this study are computed with the following specifications. All recordings in the dataset have a sampling rate of 44100 Hz. For all features the (first) frame decomposition is computed using a window size of 40 ms and hop size of 5 ms. The output of the first frame decomposition is a Mel spectrogram and a chromagram. A second frame decomposition is used to extract descriptors over 8-second windows with 0.5-second hop size. This is particularly useful for rhythmic and melodic descriptors since rhythm and melody are perceived over longer time frames. Rhythmic and melodic descriptors considered in this study are derived from the second frame decomposition with overlapping 8-second windows. Timbral and harmonic descriptors are derived from the first frame decomposition with 0.04-second windows and for consistency with rhythmic and melodic features, they are summarised by their mean and standard deviation over the second frame decomposition with overlapping 8-second windows. The window of the second frame decomposition is hereby termed as 'texture window' (Tzanetakis and Cook, 2002). The window size w of the texture window was set to 8 seconds after the parameter optimisation process described in Section 6.3.1. For all features a cutoff frequency at 8000 Hz is used since most of the older recordings do not contain higher frequencies than that. The audio content analysis process is summarised in Figure 6.3.

Rhythm and Timbre. For rhythm and timbre features a Mel spectrogram with 40 Mel bands up to 8000 Hz using Librosa (McFee et al., 2015b) is computed. To describe rhythmic content onset strength envelopes are extracted for each Mel band and rhythmic periodicities are computed using a second Fourier transform with window size of 8 seconds and hop size of 0.5 seconds. The Mellin transform is applied to achieve tempo invariance (Holzapfel et al., 2011) and rhythmic periodicities up to 960 beats per minute (bpm) are extracted. The output is averaged across low and high frequency Mel bands with cutoff at 1758 Hz. The resulting rhythmic feature vector has length 400 values. Timbral aspects are characterised by 20 MFCCs and 20 first-order delta coefficients after removing the DC component (Aucouturier et al., 2005). The mean and standard deviation of these coefficients over 8-second windows with 0.5-second hop

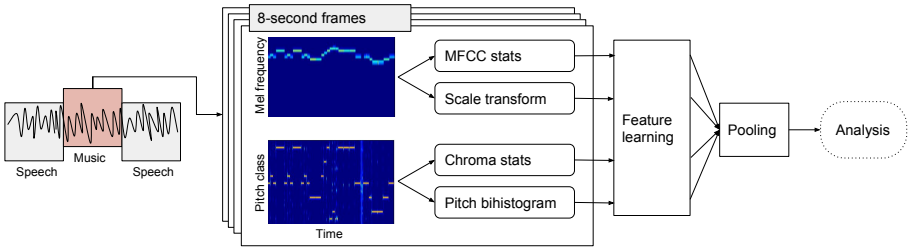


Figure 6.3: Overview of the audio content analysis process. Mel-spectrograms and chromagrams are processed in overlapping 8-second frames to extract rhythmic, timbral, harmonic, and melodic features. Feature learning is applied to the 8-second features and average pooling across time yields the representations for further analysis.

size are extracted. This results in a total of 80 feature values describing timbral aspects.

Harmony and Melody. To describe harmonic content chromagrams are computed using a variable-Q transform (Maruo et al., 2015)up to 8000 Hz with 5 ms hop size and 20-cent pitch resolution to allow for microtonality. Chromagrams are aligned to the pitch class of the maximum magnitude per recording for key invariance. Harmonic content is described by the mean and standard deviation of chroma vectors using 8-second windows with 0.5-second hop size. The dimensionality of the harmonic feature vector results in a total of 120 values. To describe melodic content pitch contours are extracted from polyphonic music signals using a method based on a time-pitch salience function (Salamon and Gómez, 2012). The pitch contours are converted to 20-cent resolution binary chroma vectors with entries of 1, whenever a pitch estimate is active at a given time, and 0 otherwise. Melodic aspects are captured via pitch bihistograms which denote counts of transitions of pitch classes (Van Balen et al., 2014). A window of $d = 0.5$ seconds is used to look for pitch class transitions in the binary chroma vectors. The resulting pitch bihistogram matrix consists of $3600 = 60 \times 60$ values corresponding to pitch transitions with 20-cent pitch resolution. For efficient storage and processing, the matrix is decomposed using non-negative matrix factorisation (Lee and Seung, 1999). Only 2 basis vectors are kept with their corresponding activations to represent melodic content. It was estimated that keeping only 2 bases was enough to provide sufficient reconstruction for most pitch bihistograms in the dataset (average reconstruction error $< 25\%$). Pitch bihistograms are also computed over 8-second windows with 0.5-second hop size. This results in a total of 120 feature values describing melodic aspects.

The scale transform implementation described above averages periodicities

across low and high frequency bands instead of averaging across all bands as described in Section 4.4 and used in corresponding publications (Panteli and Dixon, 2016; Panteli et al., 2016b). Splitting into low and high frequency bands instead of averaging across all frequencies was found to perform slightly better in preliminary experiments (Panteli et al., 2016a). The pitch bihistogram implementation described above is extracted from melodic contours (Salamon and Gómez, 2012) instead of chromagrams as described in Section 4.4 and used in corresponding publications (Panteli and Dixon, 2016; Panteli et al., 2016a,b). This is because the pitch bihistogram extracted from the chromagram was found to have significant overlap with the harmonic descriptors representing statistics of the chromagram. These modifications of the rhythmic and melodic descriptors do not affect the invariance of the features to audio transformations (Section 4.4.2).

Combining all features together results in a total of 840 descriptors for each recording in the dataset. A z-score standardisation of the 840 features is applied across all recordings before further processing.

6.2.4 Feature learning

The low-level descriptors presented above are combined with feature learning methods to learn high-level representations that best capture similarity in world music. Feature learning is also appropriate for reducing dimensionality, an essential step for the amount of data analysed in this study. Feature representations are learned from the 8-second frame-based descriptors. The country label of a recording is considered a proxy for music style and is used as ground truth for supervised training and cross-validating the methods.

There are numerous feature learning techniques to choose from in the literature. Non-linear models such as neural networks usually require large training data sets (Chen et al., 2017). Compared to these approaches a fairly limited number of audio recordings is available and the low-level descriptors partly incorporate expert knowledge of the music (Section 6.2.3). In this case, simpler feature learning techniques are more suitable for this amount and type of data. A total of 4 linear models trained in supervised and unsupervised fashions are considered.

The audio features are standardised using z-scores and aggregated to a single feature vector for each 8-second frame of a recording. Feature representations are learned using Principal Component Analysis (PCA), Non-Negative Matrix Factorisation (NMF), Semi-Supervised Non-Negative Matrix Factorisation (SS-NMF), and Linear Discriminant Analysis (LDA) methods (Sun et al., 2013). PCA and NMF are unsupervised methods and extract components that ac-

count for the most variance in the data without any prior information on the data classes. LDA is a supervised method and tries to identify attributes that account for the most variance between classes (in this case country labels). SSNMF works similarly to NMF with the difference that ground truth labels are taken into account in addition to the data matrix in the optimisation step (Lee et al., 2010).

The dataset of 8200 recordings is split into training (60%), validation (20%), and testing (20%). The models are trained and tested with the frame-based descriptors; this results in a dataset of 325435, 106632, and 107083 frames for training, validation, and testing, respectively. Frames used for training do not belong to the same recordings as frames used for testing or validation and vice versa. The training set is used to train the PCA, NMF, SSNMF, and LDA models and the validation set to optimise the parameters. The testing set is used to report the classification accuracy based on which the best projection is selected for subsequent analysis. In each experiment, feature learning components constituting to 99% of the variance are retained. Section 6.3.1 provides also an analysis of the feature weights for the components of the best performing feature learning method.

A classification task is used to assess the quality of the learned space and optimise the parameters. An ideal music similarity space separates well data points belonging to different music classes and enables good classification results to be achieved with simple classifiers. Building a powerful classifier is not a priority in this case since the aim is to assess the learned embeddings rather than optimise the classification task itself. Classifiers widely used in the machine learning community are considered (Bishop, 2006). In particular, four classifiers are trained, K-Nearest Neighbour (KNN), Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and Random Forest (RF), to predict the country label of a recording. The purpose of the classification task is to optimise the window size w of the audio descriptors and assess the quality of the learned spaces in order to select the optimal feature learning method for the data. The classification F-score metric is used to compare the performance of the models. Section 6.3.1 also provides an analysis of the coefficients of the best performing classifier.

In order to assess the contribution of different features to the classification task, five sets of features are considered: a) scale transform (rhythmic); b) MFCCs (timbral); c) average chroma vectors (harmonic); d) pitch bihistograms (melodic); and e) the combination of all the above. In each case, feature learning is applied on the selected feature set and frame-based projections are aggregated using the mean prior to classification. I also tested for aggregation using the mean and standard deviation of frame-based descriptors but this did not im-

prove results; hence it was omitted. In the case of testing the combination of all features (e), feature learning is applied for each feature set separately and the components from all feature sets are then concatenated prior to mean aggregation and classification (see also Figure 6.3). Results for the feature learning optimisation and classification experiments are presented in Section 6.3.1.

6.2.5 Outlier recordings

The feature learning and classification methods described above (Section 6.2.4) identify the optimal projection for the data. In the next step of the analysis the projected space is used to investigate music dissimilarity and identify outliers in the dataset. A recording is considered an outlier if it is distinct compared to the whole set of recordings. Outliers are detected based on a method of squared Mahalanobis distances (Aggarwal and Yu, 2001; Hodge and Austin, 2004). Using Mahalanobis, a high-dimensional feature vector is expressed as the distance to the mean of the distribution in standard deviation units. Let $X \in \mathbb{R}^{I \times J}$ denote the set of observations for I recordings and J features. The Mahalanobis distance for observation $\mathbf{x_i} = (x_1, x_2, ..., x_J)$ for recording i from the set of observations X with mean $\mu = (\mu_1, \mu_2, ..., \mu_J)$ and covariance matrix S is denoted

$$D_M(\mathbf{x_i}) = \sqrt{(\mathbf{x_i} - \mu)^T S^{-1} (\mathbf{x_i} - \mu)} \tag{6.1}$$

with $\mathbf{v}^T$ the transpose of the vector $\mathbf{v}$. Data points that lie beyond a threshold, typically set to the $q = 97.5\%$ quantile of the chi-square distribution with J degrees of freedom (Filzmoser, 2004), are considered outliers. This is denoted

$$O = \{i \in H | D_M(\mathbf{x}_i) > \sqrt{\chi^2_{J,q}}\} \tag{6.2}$$

where $H = \{1, 2, ..., I\}$ denotes the index of the observations.

Due to the high dimensionality of the feature vectors every data point can be considered far from the centre of the distribution (Filzmoser et al., 2008). To compensate for a possible large amount of outliers a higher threshold based on the $q = 99.9\%$ quantile of the chi-square distribution is considered.

To gain a better understanding of the type of outliers for each country outliers are detected using a) rhythmic, b) timbral, c) harmonic, and d) melodic features. For example, for J_R the dimensionality of the rhythmic feature vector and $X_R \in \mathbb{R}^{I \times J_R}$ the set of observations, the set of outlier recordings with respect to rhythmic characteristics is denoted

$$O_R = \{i \in H | D_M(\mathbf{x}_{R,i}) > \sqrt{\chi^2_{J_R,99.9}}\} \tag{6.3}$$

for observation $\mathbf{x}_{R,i} \in X_R$. Outliers are detected with respect to rhythmic (O_R), timbral (O_T), melodic (O_M), and harmonic (O_H) characteristics.

6.2.6 Spatial neighbourhoods

In the previous section outliers were detected by comparing a recording to all other recordings in the dataset. Here spatial relations are taken into account and recordings from a given country are compared only to recordings of its neighbouring countries. In this way spatial outliers are detected, i.e. recordings that are outliers compared to their spatial neighbours (Chen et al., 2008). Spatial neighbourhoods are constructed based on contiguity and distance criteria: a) two countries are neighbours if they share a border (a vertex or an edge of their polygon shape), b) if a country doesn't border with any other country (e.g., the country is an island) its neighbours are defined by the 3 closest countries estimated via the Euclidean distance between the geographical coordinates (latitude and longitude) of the centre of each country.

Let N_i denote the set of neighbours for country i estimated via

$$N_i = \{j \in \{1, ..., R\} | \text{j is neighbour to i}\} \tag{6.4}$$

for R the number of countries. Table A.1 provides the neighbours of each country in the dataset as estimated via this approach. The geographical boundaries of each country are derived from spatial data available via the Natural Earth platform (Kelso and Patterson).

The set of recordings from a given country is appended with recordings from neighbouring countries as defined by the country's spatial neighbourhood (Table A.1). This set is used to detect outliers with the Mahalanobis distance as defined in Equation 6.2. Spatial outliers are detected in this manner for all countries in the dataset.

6.2.7 Outlier countries

The unit of analysis in the previous sections was the individual recordings. In this section the focus is placed at the country. Outlier countries are detected in a similar manner as before where country features now summarise the information of the underlying recordings. The advantage of placing the focus at the country level is that the feature representations can now summarise the variety of styles that exist in the music of a country in the available dataset. Outliers are not judged by individual recordings but rather by the distribution of the whole set of recordings of each country.

K-means clustering is used to map recording representations to one of K

clusters. The country representation is then derived from a histogram count of the K clusters of its recordings. Let $X \in \mathbb{R}^{I \times J}$ denote the set of observations for I recordings and J features. K-means are computed for X mapping each recording to one of K clusters. A linear encoding function $f : \mathbb{R}^J \rightarrow \mathbb{R}^K$ is used so that each recording representation $\mathbf{x}_i \in \mathbb{R}^J$ for $i = 1, ..., I$ is mapped to a vector $f(\mathbf{x}_i) \in \mathbb{R}^K$ via the dot product between $\mathbf{x}_i$ and the cluster centroids $\mathbf{m}_k \in \mathbb{R}^J$ for $k = 1, ..., K$ clusters. The feature vector for a country $\mathbf{c}_r \in \mathbb{R}^K$ is the normalised histogram count of K clusters for the set I_r of recordings i from country r, denoted

$$\mathbf{c'}_r = \sum_{i \in I_r} f(\mathbf{x}_i). \tag{6.5}$$

Each histogram is normalised to the unit norm, where $\mathbf{c}_r = \frac{\mathbf{c'}_r}{|\mathbf{c'}_r|}$. Let $C \in \mathbb{R}^{R \times K}$ denote the feature representations for R countries and K clusters derived as explained above. The optimal number K of clusters is decided based on the silhouette score (Rousseeuw, 1987) after evaluating K-means for K between 10 and 30 clusters.

The similarity between countries is assessed via hierarchical clustering (Johnson, 1967). Hierarchical clustering estimates pairwise distances between observations (in this case country representations) using a distance metric d and builds a hierarchy of observations by merging or splitting sets of observations under a linkage criterion l. For consistency with the previous outlier detection method (Section 6.2.5), Mahalanobis is used as the distance metric d to estimate pairwise similarity between countries. Pairwise Mahalanobis distance between countries is denoted

$$D_M(\mathbf{c}_i, \mathbf{c}_j) = \sqrt{(\mathbf{c}_i - \mathbf{c}_j)^T \bar{S}^{-1} (\mathbf{c}_i - \mathbf{c}_j)} \tag{6.6}$$

where $\bar{S}$ is the covariance matrix and $i, j \in \{1, 2, ..., R\}$. A hierarchy of countries is constructed using the average distance between sets of observations as the linkage criterion l. This is denoted

$$l = \frac{1}{|C_I||C_J|} \sum_{\mathbf{c}_i \in C_I} \sum_{\mathbf{c}_j \in C_J} D_M(\mathbf{c}_i, \mathbf{c}_j) \tag{6.7}$$

for two sets $C_I, C_J \subset C$ of observations.

6.3 Results

Results are presented in the following order. First the parameter optimisation is presented which tunes the window size w of the audio content analysis and

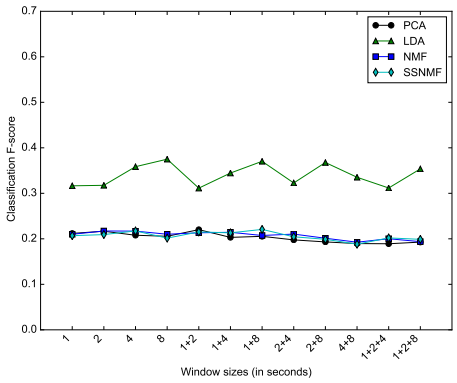


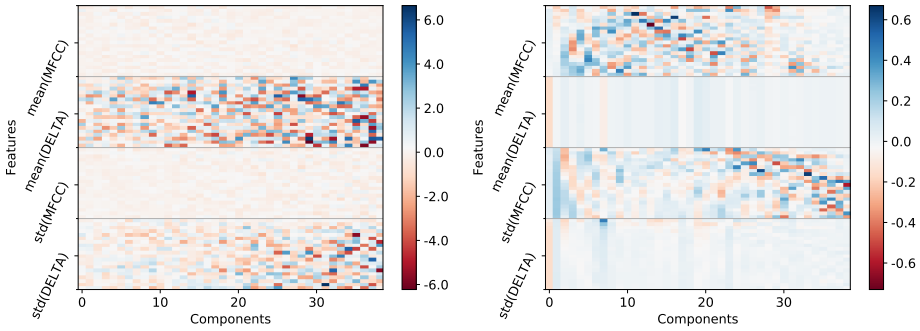
Figure 6.4: Classification F-score on the validation set for the best performing classifier (LDA) across different window sizes. Accuracies are compared for different feature learning methods (PCA, LDA, NMF, SSNMF). Combinations of window sizes are marked by '+' in (a), for example '4+8' represents the accuracy when combining features from the 4-second and the 8-second windows. Considering the performance of all feature learning methods, the optimal window size is 8 seconds.

selects the optimal feature learning method. Using the selected feature learning method outliers are detected at the global and spatial level and for different sets of features. The last part of the results presents outliers at the country level as derived via hierarchical clustering.

6.3.1 Parameter optimisation

As mentioned in Section 6.2.3, the window size w in the feature extraction process was optimised based on a classification task. Given the feature transformed representations of each recording in the training set, 4 classifiers (KNN, LDA, SVM, RF) were trained to predict the country label of a recording. Parameter optimisation was based on the classification accuracy on the validation data. Figure 6.4 shows the classification F-score of the best performing classifier (LDA) for a range of window sizes w. Based on this evaluation the optimal window size was $w = 8$ seconds with highest F-score of 0.37 for the LDA classifier in combination with the LDA-transformed features.

The dimensions of the LDA-transformed features can be explained in the following way. LDA components for the rhythmic features give more weight to the periodicities of the high-frequency Mel bands (above 1758 Hz). Melodic features receive similar weights for both the bases and activations of the pitch bihistogram. LDA components for the harmonic features assign more weight to relative pitch values (mean of chroma vectors) rather than pitch fluctuations (standard deviation of chroma vectors) over time. LDA components for timbral



(a) LDA (b) PCA

Figure 6.5: LDA and PCA components weigh timbral features in opposite ways. LDA components focus on timbre fluctuation (mean and standard deviation of MFCC delta coefficients) over time whereas PCA components focus on absolute timbre qualities (mean and standard deviation of MFCC coefficients) over time.

features focus on timbre fluctuation (mean and standard deviation of MFCC delta coefficients) over time. This is opposite to the behaviour of the PCA transformation where components focus on absolute timbre qualities (mean and standard deviation of MFCC coefficients) over time. Figure 6.5 illustrates the difference between LDA and PCA components for the timbral features.

6.3.2 Classification

The classification results for the different classifiers in combination with the feature learning methods are presented in Table 6.1. Only the best performing classifier is discussed further in the text for simplicity. Classification accuracy of the test set was assessed after fixing the window size of the feature extraction to $w = 8$ seconds as found optimal in Section 6.3.1. Results suggest that the best classifier when the combination of all features is considered is the LDA classifier with the LDA-transformed features (classification F-score of 0.321). Rhythmic, melodic, and harmonic features achieved best classification performance for the LDA-transformed features and the LDA classifier whereas timbral features achieved best classification performance for the LDA-transformed features and the SVM classifier. The first 10 components of the LDA classifier trained with the LDA-transformed features (for the combination of 'all' features) give more weight to the timbral and harmonic dimensions and explain 24% of the variance. The remaining components give more weight to the rhythmic and melodic dimensions. More information on the classification results and confusion matrices can be found in the published code repository (`http://github.com/mpanteli/music-outliers`).

Transform	Classifier	F-measure				
		All	Rhythm	Melody	Timbre	Harmony
LDA	LDA	**0.321**	**0.150**	**0.070**	0.199	**0.107**
SSNMF	LDA	0.183	0.053	0.039	0.165	0.082
NMF	LDA	0.178	0.059	0.046	0.166	0.086
–	LDA	0.177	0.060	0.038	0.191	0.084
PCA	LDA	0.175	0.055	0.046	0.162	0.084
LDA	KNN	0.152	0.055	0.023	0.282	0.086
SSNMF	KNN	0.143	0.043	0.015	0.227	0.072
PCA	KNN	0.141	0.053	0.027	0.221	0.081
–	KNN	0.140	0.052	0.027	0.222	0.082
NMF	KNN	0.114	0.043	0.029	0.178	0.080
–	RF	0.083	0.040	0.032	0.114	0.057
LDA	RF	0.071	0.031	0.017	0.150	0.051
NMF	RF	0.063	0.032	0.020	0.126	0.042
PCA	RF	0.046	0.026	0.019	0.140	0.045
SSNMF	RF	0.045	0.031	0.018	0.116	0.035
LDA	SVM	0.023	0.079	0.050	**0.296**	0.090
SSNMF	SVM	0.021	0.011	0.005	0.019	0.014
NMF	SVM	0.016	0.008	0.008	0.011	0.012
–	SVM	0.015	0.047	0.038	0.250	0.088
PCA	SVM	0.015	0.048	0.039	0.246	0.092

Table 6.1: Classification F-scores of the test set for the country of recording (– denotes no transformation). The window size of the features is 8 seconds as found optimal in section Parameter optimisation. Results are sorted by highest to lowest F-score of the combination of all features ('All').

6.3.3 Outliers at the recording level

The classification results (Section 6.3.2) indicated the optimal feature learning method (LDA) that best approximates music similarity in this dataset. The LDA-projected space was used to investigate music dissimilarity and identify outliers in the dataset.

From a total number of 8200 recordings, 1706 recordings were detected as outliers. The distribution of outliers per country, normalised by the number of recordings per country in the dataset, is summarised in Figure 6.6. The country with the most outliers is Botswana with 61% (55 out of 90) of its recordings identified as outliers, followed by Ivory Coast (60%, 9 out of 15), Chad (55%, 6 out of 11), and Benin (54%, 14 out of 26). The percentage of outliers per country was not significantly correlated with the number of recordings sampled from that country (Pearson correlation coefficient $r = -0.01$ with p-value=0.91).

Listening to some examples the following timbral characteristics can be summarised for the outlier recordings. Outliers from Botswana include solo performances of the mouthbow and dance songs featuring group singing accompanied with handclapping or other percussion. Outlier recordings from Ivory Coast feature music from the Kroo ethnic group who originated in eastern Liberia and consist of singing melodies accompanied by woodwind instruments and guitar.

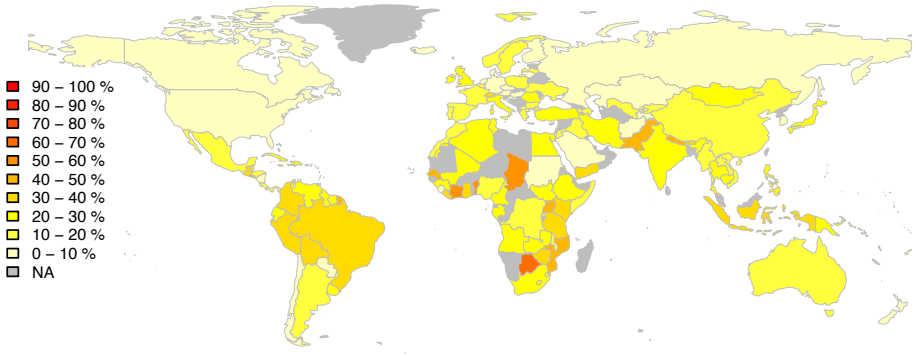
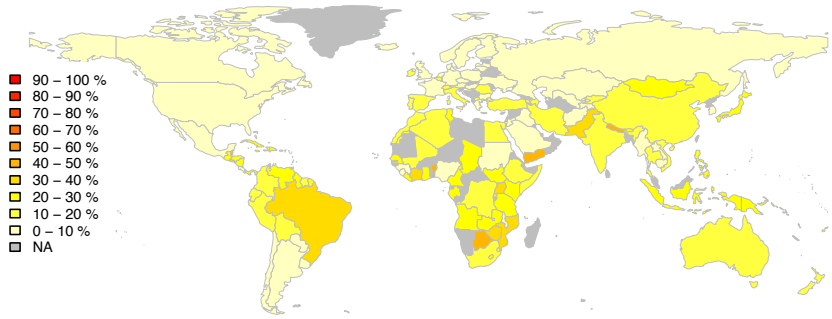


Figure 6.6: Distribution of outliers per country. The colour scale corresponds to the normalised number of outliers per country, where 0% indicates that none of the recordings of the country were identified as outliers and 100% indicates that all of the recordings of the country are outliers.

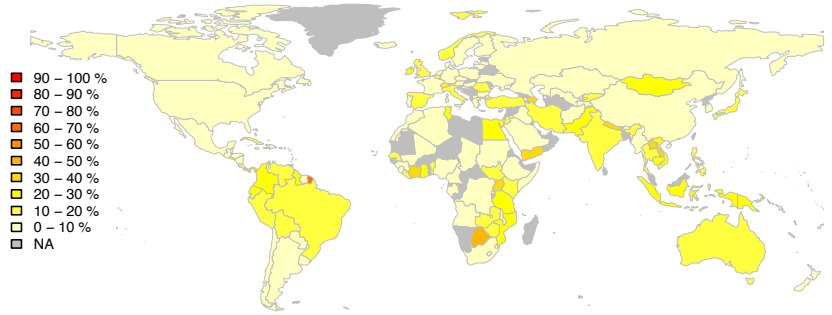
Outlier recordings from Chad feature mainly dance music with emphasis on percussive and wind instruments as well as examples of the singing voice in solo and group performances. Outlier recordings from Benin include solo performances of the Yoruba drums and music from the Fon culture including examples of group singing with gong accompaniment.

To gain a deeper understanding of the type of outliers for each country, outliers using a) rhythmic, b) timbral, c) melodic, and d) harmonic features were detected. Results are shown in Figure 6.7. With respect to rhythmic aspects the countries with the most outliers are Benin (50%, 13 out of 26), Botswana (49%, 44 out of 90), and Nepal (42%, 40 out of 95). The countries with the most outliers with respect to timbral characteristics are French Guiana (78%, 19 out of 28), Botswana (48%, 43 out of 90), and Ivory Coast (40%, 5 out of 13). The countries with the most outliers with respect to melodic aspects are Zimbabwe (53%, 8 out of 15), Uruguay (48%, 15 out of 31), and Guinea (46%, 5 out of 11) and with respect to harmonic aspects Benin (54%, 14 out of 26), Pakistan (46%, 42 out of 91), and Gambia (36%, 18 out of 50).

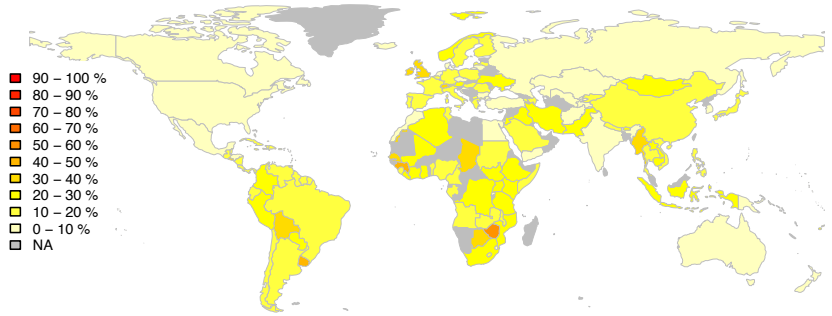
Listening to some examples the following characteristics can be summarised for these outliers. Rhythmic outliers include examples from African polyrhythms as well as examples with frequent transitions between binary and ternary subdivisions. The most prominent instruments in the rhythmic outliers are pitched and non-pitched percussion. Most rhythmic outliers tend to have a high density of events, i.e., there are many onsets within each bar duration. Outliers with respect to timbral characteristics include solo performances of xylophones and gongs for example recordings from Botswana, Indonesia, and Gamelan recordings from the Philippines. Another category of instruments that often gives rise to timbre outliers are wind instruments such as reedpipes and flutes. Outliers



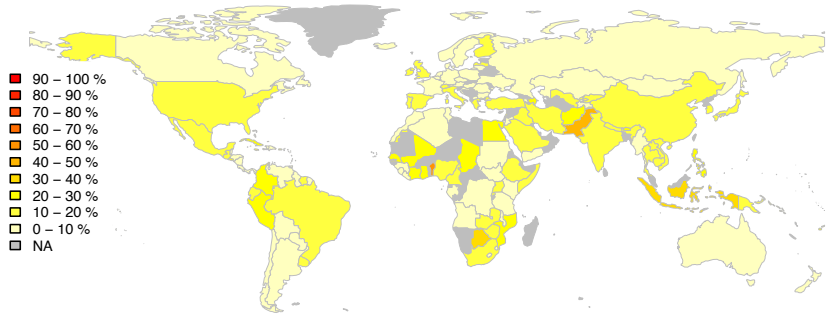
(a) Rhythm



(b) Timbre



(c) Melody



(d) Harmony

Figure 6.7: Distribution of outliers per country for each feature type. The colour scale corresponds to the normalised number of outliers per country, from 0% of outliers (light colours) to 100% (dark colours).

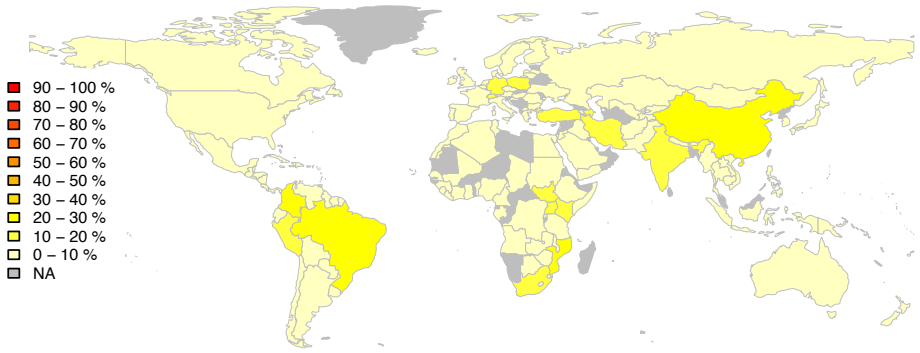


Figure 6.8: Distribution of outliers per country for the spatial neighbourhoods shown in Table A.1. The colour scale corresponds to the normalised number of outliers per country, from 0% of outliers (light colours) to 100% (dark colours).

with respect to melodic characteristics include polyphonic melodies performed on the accordion (e.g., recordings from Uruguay) or the mbira (e.g., recordings from Zimbabwe). With respect to harmony, outliers exhibit microtonal scales and feature instruments with distinct tuning, for example solo sitar or surnai performances from Pakistan, xylophone and gong performances from Benin and Indonesia. Listening examples can be found at the online demo (`http://mpanteli.github.io/music-outliers/demo/outliers`).

Spatial outliers

In the previous section outliers were detected by comparing a recording to all other recordings in the dataset. In this section spatial relations are taken into account and recordings from a given country are compared only to recordings from the neighbouring countries (Section 6.2.6). Figure 6.8 provides an overview of the distribution of spatial outliers, normalised by the total number of recordings in each spatial neighbourhood. Results show that China is the country with the most spatial outliers (26%, 26 out of 100), followed by Brazil (24%, 24 out of 100), Colombia (21%, 19 out of 90), and Mozambique (21%, 7 out of 34).

China is the country with most spatial neighbours, bordering with 12 other countries in the dataset (Table A.1). Recordings from China feature the butterfly harp string instrument and singing examples from the Han cultural group, often with a bright sound and prominent singing in relatively high frequencies. These examples are compared to various instruments and music styles from the neighbouring countries including lute performances from Kyrgyzstan, Mongolian Jewish harp, Indian tala, Nepalese percussion and wind instrument performances, polyphonic singing from Vietnam and Laos, and instrumental pieces featuring the balalaika from Russia. Compared to the analysis of global

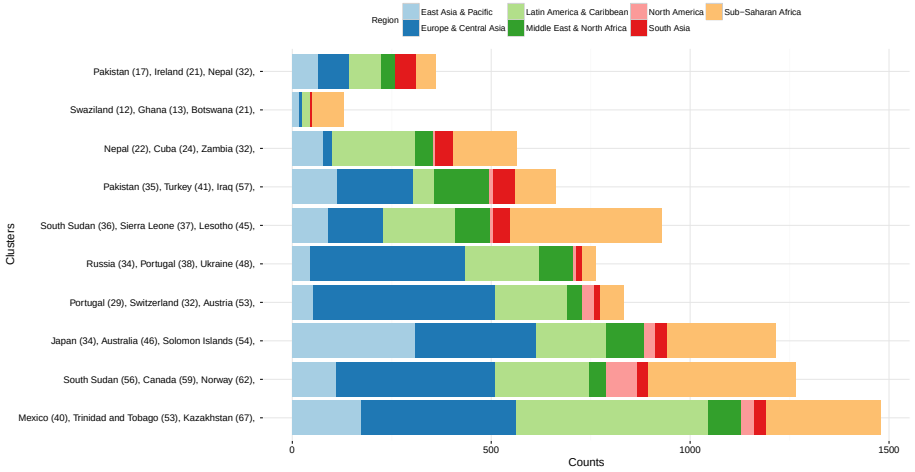


Figure 6.9: The 3 most frequent countries and their corresponding number of recordings for each of the 10 clusters.

outliers (Figure 6.6) recordings from China stand out only with respect to its spatial neighbourhoods but are not so distinct compared to the whole dataset of world music.

6.3.4 Outliers at the country level

In this section the focus is placed on the country instead of the individual recordings as the unit of analysis and outlier countries are detected as described in Section 6.2.7. The silhouette score indicated an optimal number of $K = 10$ clusters. The 3 most frequent countries in each cluster are shown in Figure 6.9.

The similarity between countries was estimated via hierarchical clustering. Results are presented in a dendrogram in Figure 6.10. The countries with the most distinct feature representations are South Sudan, Botswana, Ghana, Austria and Switzerland (in order of most to least distinct). The aforementioned countries were found dissimilar (with respect to a threshold) to any other country in the dataset of 137 countries.

Recordings from South Sudan feature mostly examples of the singing voice in solo and group performances. The use of solely the singing voice is what likely makes the feature representation of South Sudan so different from other countries. Solo and group singing examples occur in many other countries but the consistent use of these examples in recordings from South Sudan makes its music style distinct. A similar observation holds for recordings from Austria and Switzerland featuring mostly dance songs with accordion accompaniment. This might not be a unique music style across the whole dataset but the consistent use

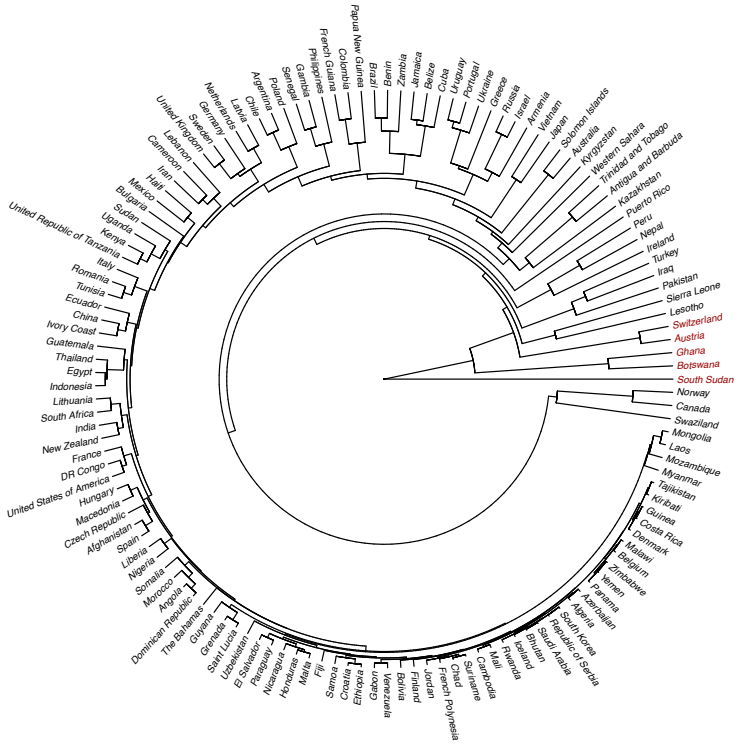


Figure 6.10: Hierarchical clustering of the 137 countries in the dataset. Countries, represented as nodes in the dendrogram, that lie in the same branch are similar (the shorter the branch the higher the similarity). Each branch denotes a cluster and pairs of clusters are merged as one moves up the hierarchy. The most distinct countries are annotated with red colour.

of this style in the recordings from Austria and Switzerland makes their music distinct from other countries. Botswana and Ghana, also detected as outlier countries with the hierarchical clustering approach, exhibit the use of a variety of music styles. Botswana was also detected as the country with the most outlier recordings compared to the global dataset (Section 6.3.3). Figure 6.10 also revealed some music similarity relationships between countries of geographical or cultural proximity. However, as the scope of this study is rather on music dissimilarity and outliers the exploration of these relationships is left for future work.

6.4 Subjective evaluation

The methodology described above combining audio content analysis with outlier detection to find music examples that stand out from the corpus was also evaluated qualitatively (Panteli et al., 2016a). In particular, a listening experiment

in the 'odd one out' fashion (Wolff and Weyde, 2011) was performed. Listeners were asked to evaluate triads of audio excerpts by selecting the one that is most different from the other two, in terms of its musical characteristics. For the purpose of evaluating outliers, a triad consisted of one outlier excerpt and two non-outliers as estimated by their Mahalanobis distance from the set of all recordings.

Recordings that are close to the mean of the distribution, i.e., they denote the most typical examples, are hereafter called 'inliers'. To distinguish outliers from inliers and other excerpts which are neither outliers nor inliers, an upper and a lower threshold are set for the Mahalanobis distance. Distances above the upper threshold identify outliers, and distances below the lower threshold identify inliers. The thresholds are selected such that the majority of excerpts are neither outliers nor inliers. A random sample of 60 outliers is selected given these thresholds. For each of these outliers five pairs of inliers are selected at random, resulting in a total of 300 triads (5 triads for each of 60 outliers), which are split into 10 sets of 30 triads.

Each participant rates one randomly selected set of 30 triads. The triads are presented in random order to the participant. Additionally 2 control triads are included to assess the reliability of the participant. A control triad consists of two audio excerpts (the inliers) extracted from the first and second half, respectively, of the same recording and exhibiting very similar musical attributes, and one excerpt (the outlier) from a different recording exhibiting very different musical attributes. At the end of the experiment a questionnaire is attached for demographic purposes.

A total of 23 subjects participated in the experiment. There were 15 male and 8 female participants and the majority (83%) were aged between 26 and 35 years old. A small number of participants (5) reported they were very familiar with world music genres and a similar number (6) reported they were quite familiar. The remaining participants reported they were not so familiar (10 of 23) and not at all familiar (2) with world music genres.

Following the specifications described above, each participant's reliability was assessed with two control triads and results showed that all participants rated both of these triads correctly. From the data collected, each of the 300 triads (5 triads for each of 60 detected outliers) was rated a minimum of 1 and maximum of 5 times. Each of the 60 outliers was rated a minimum of 9 and maximum of 14 times with an average of 11.5.

A total of 690 ratings (23 participants rating 30 triads each) were obtained. For each rating an accuracy value of 1 was assigned if the odd sample selected by the participant matched the 'outlier' detected by the algorithm versus the two 'inliers' of the triad, and an accuracy of 0 otherwise. The average accuracy

from 690 ratings was 0.53. A second measure aimed to evaluate the accuracy per outlier. For this, the 690 ratings were grouped per outlier, and an average accuracy was estimated for each outlier. Results showed that each outlier achieved an average accuracy of 0.54 with standard deviation of 0.25. One particular outlier was never rated as the odd one by the participants (average accuracy of 0 from a total of 14 ratings). Conversely, four outliers were always in agreement with the subjects' ratings (average accuracy of 1 for about 10 ratings for each outlier). Overall, there was agreement well above the random baseline of 33% between the automatic outlier detection and the odd one out selections made by the participants.

According to the outcome of the listening test the computational framework proposed above detects music outliers with accuracy moderately agreeing with that of human perception. Participants of the listening test commented also on the difficulty of the task due to the music segments being very different from each other. For example, some triads consisted of music segments that were different in e.g. both their rhythm and timbre which made it difficult to judge the odd one out. Limiting the diversity of music styles and focusing on specific musical attributes for such a comparison could improve the reliability of the ratings. Further analyses such as how the music culture and music education of the participant influences the similarity ratings and which examples are more challenging for computational versus human annotation are left for future work.

6.5 Discussion

World music recordings from two large archives were combined and a methodology to extract music features and detect outliers in the dataset was proposed. Findings explored differences across world music and revealed geographical patterns of music outliers.

Several pre-processing steps were taken into account to isolate relevant music information from the audio signal: speech segments were separated from music, frequencies above 8000 Hz were omitted for consistency with old recording equipment, and low-level music descriptors were combined with feature learning to give higher-level representations respecting the world music characteristics. The size of the texture window was optimised and longer windows (8 seconds) provided better representations of the music data than shorter ones ($4, 2, 1$ seconds). Feature learning was better in the supervised setting (LDA outperformed PCA and NMF) even though class labels (in this case countries) were not necessarily unique identifiers for the underlying musical content.

A method to detect outliers was proposed and several ways of understanding the musical differences were explored. The countries with the most outlier

recordings were listed and the analysis was expanded to explain which music features were distinct in these outliers. For example, Botswana was the country with most of its recordings detected as outliers and feature analysis showed that these outliers were mostly due to rhythmic and timbral features. With respect to rhythmic features, African countries indicated the largest amount of outliers with recordings often featuring the use of polyrhythms. Harmonic outliers originated mostly from Southeast Asian countries such as Pakistan and Indonesia, and African countries such as Benin and Gambia with recordings often featuring inharmonic instruments such as the gong and bell.

A sensitivity experiment was implemented to check how stable the outlier findings are with respect to different datasets. The outlier analysis was repeated 10 times, each time selecting at random a stratified sample of 80% of the original dataset. That is, outliers were detected for subsets of around 6560 recordings of the whole dataset of 8200 recordings. The distribution of outliers per country, especially for countries detected with the most outliers, is expected to be similar across all 10 repeated experiments. The majority vote of outlier countries resulting in the top $K = 10$ positions of each experiment was used as the ground truth. Assessing the precision at $K = 10$ for each experiment assuming majority vote ground truth showed that the geographical patterns of outliers (Figure 6.6) were on average consistent across multiple random subsets of the original dataset (precision at K mean $= 0.67$, standard deviation $= 0.06$).

Recordings from neighbouring countries were compared by incorporating spatial information. This gave rise to music cultures that are not distinct compared to the global dataset but are still unique compared to their spatial neighbours. For example, music from China with bright timbres was found to be unique compared to its many spatial neighbours. Music from Brazil was also distinct compared to its spatial neighbours, an observation that could be attributed to cultural differences the most obvious of which is the use of different languages between Brazil and its neighbouring countries. Proving historical and cultural influence was not the aim of this study but these findings could provide a good starting point for further investigation.

A method to extract feature summaries for each country was also proposed and clusters were estimated for the whole set of recordings. A total of 10 clusters was found to best represent the music styles in the dataset and recordings from geographically similar regions often clustered together. Hierarchical clustering at the country level representation revealed African countries such as South Sudan, Botswana, and Ghana as most distinct from others in the dataset.

6.5.1 Hubness

This research deals with high dimensional vectors (features of 840 dimensions as described in Section 6.2.3) and analysis of nearest neighbour relationships. High dimensional spaces are prone to produce data points that appear in the neighbourhood of other points disproportionately often. This could also be one of the reasons that too many outliers were detected in the dataset (almost 20% from the total of 8200 recordings were detected as outliers.) The effect of hubness in the data was tested following the approach suggested by Schnitzer et al. (2012). Hubness was estimated as the skewness of the n-occurrence where n-occurrence defines the number of times a given track x occurs in the top n neighbours of other tracks. Pairwise Mahalanobis distances were used and the n nearest neighbours for each track in the dataset were assessed for $n = 60$, the average number of recordings per country. A positively skewed distribution with hubness = 10.1 was observed. A total of 129 out of 8200 recordings occurred in the nearest neighbour lists of more than 1000 tracks (2% large hubs) and 3332 recordings had n-occurrence = 0 (41% orphans). Pairwise Mahalanobis distances in this study are only used for the computation of outlier countries (Section 6.3.4). Future work could aim to reduce hubness via local scaling or mutual proximity (Schnitzer et al., 2012).

6.5.2 Future work

There are several steps in the overall methodology that could be implemented differently and audio excerpts and features could be expanded and improved. Numerous audio features have been proposed in the literature for describing musical content in sound recordings for various applications. A small set of features from the MIR domain was selected based on their state-of-the-art performance and relevance for world music analysis. It is clear that any such set of features does not capture all aspects of a set of musical recordings. Future work could explore the suitability of feature sets proposed by ethnomusicologists (Savage et al., 2012) or embeddings learned from raw audio or spectrograms (Dieleman and Schrauwen, 2013). The latter is partly explored in the next chapter.

In this study country labels were considered a proxy to music style and used to train models for music similarity and dissimilarity. While countries provide a broad notion of ethnic boundaries, music styles are not homogeneous within these boundaries. A country may exhibit several music styles and a music style may spread across many countries. The ambiguity of these boundaries provides an upper limit to the performance of the proposed models. This ambiguity could be reduced by incorporating more information, for example the culture or language of the musicians, to better approximate the music style of a recording.

Language information is considered in the music similarity model of Chapter 7.

This study focused on the detection of outliers in music collections. The music descriptors were derived via a multi-step procedure of processing the audio signal. The suitability of the audio tools can be questioned with regard to their ability to capture and represent high-level musical concepts (Fink, 2013). Likewise, the patterns observed could sometimes be artifacts of the computational tools. Quantitative and qualitative evaluation could be expanded to measure effects from recording date differences or acoustic environments as described in related research (Flexer and Schnitzer, 2010; Sturm, 2014, 2016).

6.6 Outlook

In this chapter a model to study music dissimilarity in a large world music corpus was developed. Low-level descriptors of rhythm, melody, timbre, and harmony were combined with machine learning methods to extract high-level world music representations. Linear feature learning techniques were tested and the supervised LDA projection performed best according to a classification task. The LDA projections were used for subsequent analysis of music dissimilarity and outlier detection. Several analyses were performed to explore the geographical patterns of music outliers. Firstly, a comparison using the combination of all features revealed recordings that stand out from the whole set of recordings. Secondly, considering each feature individually, outliers revealed geographical areas with distinct uses of rhythm, melody, timbre, and harmony. Finally, the combination of all features was considered to reveal recordings that stand out when compared to recordings from (only) their neighbouring countries. The proposed findings revealed geographical regions that seemed to maintain unique musical characteristics providing a good starting point for further investigation into the history of musical exchange.

The methodology of this chapter used a combination of features extracted with expert knowledge and machine learning methods. This was considered an improved method over the custom feature design approach presented in Chapter 5. A range of linear feature learning methods were explored and the performance of the optimal transformation method was well above the random baseline (classification F-score for the best performing feature learning method for 137 countries was 0.321 compared to the random baseline of approximately 0.010). Results could however be improved with the application of non-linear models as well as the expansion of the training set.

Non-linear projection methods achieved state-of-the-art results in related music classification and retrieval tasks (Choi et al., 2017a). In the next chapter non-linear methods learning audio features directly from the world music data

are explored. The developments in the next chapter complement the ones in the current chapter by applying a data-driven methodology, learning similarity relations from richer metadata including spatial, temporal and cultural information, and placing the focus of the application on exploring music similarity instead of dissimilarity.

Chapter 7

A study on unexpectedly similar music

In Chapter 6 a model for music dissimilarity was developed combining low-level MIR descriptors with machine learning methods. In this chapter, audio features are learned directly from the Mel spectrograms with deep neural networks and music similarity is modelled by optimising a music tag prediction task. The audio content similarity of recordings is computed from the learned embeddings. Geographical and cultural metadata are used to approximate similarities between recordings. Results focus on finding pairs of recordings that appear to have similar audio content but share no metadata. These pairs are referred to as 'unexpectedly similar' and analyses show geographical patterns of these unexpected similarities.

7.1 Motivation

In this chapter the focus is placed on a model of music similarity. Automatic systems for modelling music similarity are usually trained on a ground truth of music similarity annotations, created manually by human listeners (Wolff and Weyde, 2011; Prockup et al., 2015; Panteli et al., 2016c) or inferred from available metadata (Tzanetakis and Cook, 2002; Pampalk et al., 2005; Seyerlehner et al., 2014; Choi et al., 2016). For example, in the latter approach algorithms can be trained to predict music as similar whenever music tracks share tags of genre, mood, or artist. The approach in this chapter follows that of an automatic music tagging system, i.e., a multi-label classification system that predicts music tags from the audio content of recordings. The assumption here is that if the predictions of music tags are accurate enough, the model has learned to

map audio content to a musically meaningful space. The learned space is used for further exploration of music similarity.

Several approaches have explored the properties of music tagging, including music genre classification, to learn relationships of music similarity. These span a range of hand-crafted features (Pampalk et al., 2005; Barbedo and Lopes, 2008; Seyerlehner et al., 2014), and more recently, deep learning methods (Dieleman and Schrauwen, 2014; Nam et al., 2015; Choi et al., 2016). Compared to the methodology in Chapters 5 and 6, this chapter aims to explore music similarity relationships with a data-driven approach. The music tagging approaches reviewed below focus on studies using deep neural networks.

Convolutional neural networks (CNNs) are popular and powerful machine-learning models most used in computer vision (Krizhevsky et al., 2012). In MIR, they have been succesfully used for the task of music tagging trained on audio spectrograms (Dieleman and Schrauwen, 2014), Mel spectrograms (Choi et al., 2016), or constant-Q transforms (Oramas et al., 2017). Choi et al. (2016) trained a CNN from Mel spectrograms to predict tags of mood, genre, and era for recordings from the Magnatagatune dataset. The learned embeddings were used in other classification and retrieval tasks demonstrating that the network had learned musically relevant attributes of the signal (Choi et al., 2017a). Oramas et al. (2017) trained their CNN from constant-Q transforms for a multi-label music genre classification task. Different architectures were explored assessing the size and number of convolutional filters in each layer, and results were compared across different modalities combining audio, text, and image data.

Deep learning methods have also been explored in other MIR and audio content description tasks. A CNN approach trained on harmonic constant Q-transforms achieved state-of-the-art results in estimating multi-f_0 salience representations (Bittner et al., 2017). A related approach with CNNs explored the mapping from spectrograms to salience representations for the task of singing voice detection (Schlüter, 2016). Temporal features were also learned with a CNN trained on Mel spectrograms with filter shapes adjusted based on musical properties of rhythm and tempo (Pons and Serra, 2017).

The aforementioned approaches focused on studying properties of (mainly) Western music corpora. The analysis of world music corpora imposes further challenges, amongst others, the degraded audio quality as also discussed in Section 4.4 and, with respect to the metadata, the inconsistent taxonomy of world music genres and tags (Serra, 2011). As seen in Section 3.1, the metadata in world music collections is usually limited to spatio-temporal information such as the country and year of the recording and cultural background of the performers, for example the language and ethnic group or culture. These are different from the tags encountered in Western music such as the mood and artist. Pre-

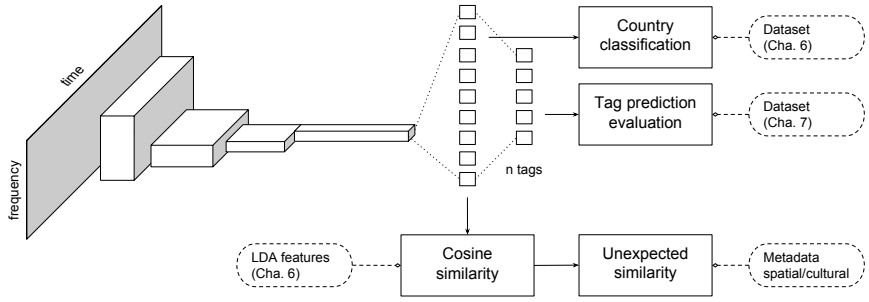


Figure 7.1: Overview of the methodology for the development of a music similarity model and the study of unexpectedly similar music.

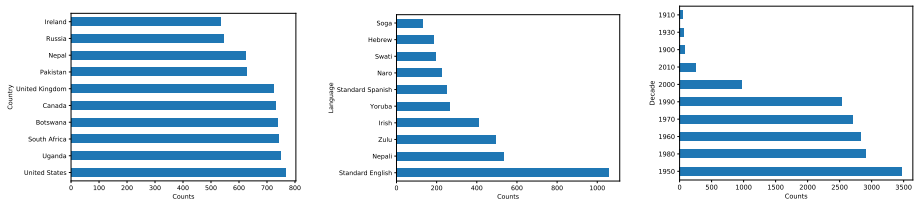
trained models for automatic Western music tagging are not suitable for world music tagging therefore a new system is developed for this purpose.

Given the state-of-the-art and challenges mentioned above a deep convolutional neural network is trained for music similarity in world music. The network learns mappings from Mel spectrograms to high-level embeddings optimised for the task of music tagging. Content-based similarity between recordings is computed by the cosine distance of the learned embeddings. The music similarities modelled from solely the audio content are compared to similarities derived from a ground truth, in this case from the metadata of recordings.

The ground truth, or alternatively referred to as the 'music similarity expectations', represents geographical and cultural aspects captured in the metadata of each recording. For example, two recordings can be expected to be similar if they come from the same or neighbouring countries, or share the same language. The goal of this study is to uncover hidden connections between music cultures. That is, identify music recordings that are similar in their audio content but share no other metadata, i.e., they are 'unexpectedly similar'. In this way, musically similar recordings could reveal possibly unknown patterns of musical exchange in the world.

7.2 Methodology

The methodology for the development of a music similarity model and its application to uncover unexpected similarities is summarised as follows. A convolutional neural network is trained from the Mel spectrograms of recordings to predict various music tags. Optimising the music tag prediction yields a mapping from low-level signal representations to a high-level latent space in which music similarity relations can be explored. The space is evaluated by the accuracy in predicting music tags as well as the accuracy of the country classification



(a) Country (b) Language (c) Decade

Figure 7.2: The 10 most frequent (a) countries, (b) languages, and (c) decades in the dataset of 18054 recordings used to study music similarity.

task presented in Chapter 6, Section 6.3.2.

The penultimate layer of the network represents the high-level audio features that can be used to study music similarity. These features are combined with the LDA features derived in Chapter 6 for optimal performance. This results in a combined feature vector which includes content-based features derived from the CNN model trained on country, language, and decade tags, and features derived from the LDA model trained on a set of country labels. The cosine distance is used to compute similarities between the audio feature vectors of each recording. The nearest $K = 5$ neighbours for each recording are considered and audio content similarity is contrasted with similarity derived from the metadata. Metadata like the country and language of a recording denote the music similarity expectations and recordings that are similar in their audio content but share no metadata are considered to be unexpectedly similar. Statistics are derived from the set of unexpectedly similar recordings and observations summarise the unexpected links between the music of different countries. Figure 7.1 provides an overview of the methodology and more details for each step are provided in the sections below.

7.2.1 Dataset

The dataset for studying music similarity is sampled from the BLSF corpus with two main criteria: a) A maximum of 1000 recordings per country (selected at random) is set to avoid the over-representation of the popular countries. b) A minimum of 50 recordings per tag (selected at random) is set to ensure enough training examples for the automatic music tagging system. Tags from 3 broad categories were considered, namely, country, language, and decade. Combining spatial, cultural, and temporal information to define the music style of a recording is considered an advantage over the previous developments using solely country information (Chapter 6) as also discussed in Section 6.5.2.

The final dataset consisted of 18054 recordings from a total of 80 countries,

33 languages, and 11 decades (124 tags in total). The 10 most frequent countries, languages, and decades in the dataset are shown in Figure 7.2. Overall, the dataset included a large representation (more than 500 recordings) of the countries Botswana, Canada, South Africa, Uganda, United States of America, United Kingdom, and a large representation (more than 400 recordings) of the languages English, Nepali, Zulu, and Irish. The recordings spanned the decades between $1900 - 2010$ with small representation for the decades before 1950 and after 2000. The list of recordings and associated metadata in this dataset can be found at `http://github.com/mpanteli/phd_thesis/blob/master/chapter_7/chapter_7_dataset_metadata.csv`.

In addition to this dataset, the set of 8200 recordings used in Chapter 6 is also considered for purposes of model evaluation and comparison. In particular, the features derived from the proposed deep learning methods are compared to the features derived from the linear feature learning methods described in Chapter 6. A classification task predicting the country of the recording (Section 6.2.4) is used for comparing the performance of the different features. The dataset of 8200 recordings has a 54% overlap with the dataset of 18054 recordings. Instances in the test set used for the classification task that are found to overlap with instances in the training set of the deep learning model are removed. This results in a smaller test set (933 recordings from a total of 134 countries) than the one reported in Section 6.3.2 (1640 recordings from a total of 137 countries). Unexpected similarities in this chapter are explored over the dataset of 8200 recordings because this dataset provides a wider geographical spread of world music than the dataset of 18054 recordings (137 countries in the former but only 80 countries in the latter). Although the latter dataset is larger, and consists of more training samples per target label which is necessary for training a deep neural network, it is not as diverse.

7.2.2 Convolutional neural networks

Following state-of-the-art research in music tagging and similarity (Choi et al., 2016, 2017a), a CNN trained on Mel spectrograms is considered. Music tagging is treated as a multi-label classification problem where multiple target labels may be assigned to each classifiable instance. The target labels are often encoded as high-dimensional sparse binary vectors but low-dimensional embeddings have been explored as well (Oramas et al., 2017). A common architecture for music tagging with CNNs consists of multiple 2D convolutional layers followed by dense layers, the last of which corresponds to logistic regression estimating the probability for each tag with the sigmoid function (Choi et al., 2016). Figure 7.3 depicts a common CNN architecture for music tagging.

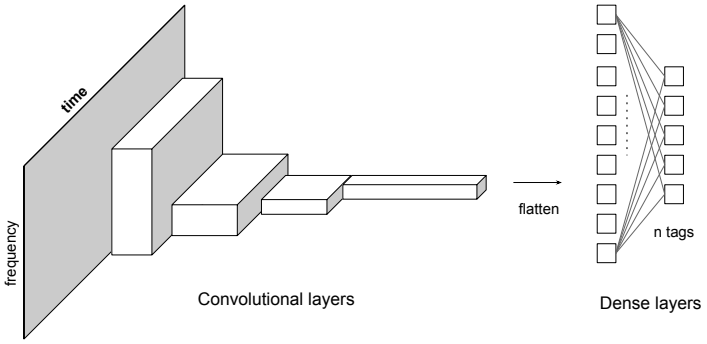


Figure 7.3: A common CNN architecture for a music tagging system.

In this chapter two architectures are considered for the task of music tagging inspired by state-of-the-art research. a) A CNN of two 2D convolutional layers (denoted CNN-2L) with filter shapes (5×5) followed by one dense layer is considered. This architecture aims to test a small network (only 2 convolutional layers and 1 dense layer) and is expected to be fairly robust to overfitting given that the world music dataset studied here is relatively small. b) A CNN with four 2D convolutional layers (denoted CNN-4L) with filter shapes (3×3) followed by two dense layers is considered. This architecture is similar to the best performing CNN model proposed by Choi et al. (2016), but filter shapes and pooling parameters are adjusted to match the dimensions of the current Mel spectrograms (Section 4.1.1). The final layer in each model is a dense layer that outputs predictions for each target label with the sigmoid activation function. Max pooling over 2D patches, batch normalisation, and leaky rectifier activations are applied in each layer whereas dropout is applied to the dense layers only (Schlüter, 2016). The models are trained with cross-entropy loss and Adam optimisation. The architectures of the two models are summarised in Table 7.1.

7.2.3 Model evaluation

The models are trained using a training and validation set and their performance is assessed on a separate testing set. For this purpose the dataset of 18054 recordings is split into train (60%), validation (24%), and test (16%) sets with no overlapping recordings between the three subsets.

Music tagging predictions are evaluated with the area under the Receiver Operating Characteristic (ROC) curve metric, referred to as AUC hereafter. The ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold changes. The AUC can be interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example.

CNN-2L		CNN-4L	
Layer	Parameters	Layer	Parameters
Convolution 2D	64 filters (5×5)	Convolution 2D	32 filters (3×3)
Max Pooling 2D	5 freq $\times$ 5 time	Max Pooling 2D	2 freq $\times$ 2 time
Convolution 2D	64 filters (5×5)	Convolution 2D	32 filters (3×3)
Max Pooling 2D	all freq $\times$ all time	Max Pooling 2D	2 freq $\times$ 2 time
		Convolution 2D	64 filters (3×3)
		Max Pooling 2D	2 freq $\times$ 2 time
		Convolution 2D	128 filters (3×3)
		Max Pooling 2D	all freq $\times$ all time
Dense	256 dimensions	Dense	128 dimensions
		Dense	64 dimensions
Dense	124 tags	Dense	124 tags

Table 7.1: The architecture of the two CNN models considered for the task of world music tagging.

This is a common metric in the evaluation of multi-label classification systems (Dieleman and Schrauwen, 2013; Choi et al., 2016; Oramas et al., 2017).

In addition to the AUC metric, the recall at $K = 10$, denoted 'Recall@K', is also computed. Recall@K denotes the proportion of relevant items found in the top K predictions. More specifically, the music tagging system outputs a probability-like estimate for each of the 124 tags in the dataset and the top 10 predicted tags are retained for the calculation of Recall@K. Each recording is represented by a country, language, or decade so there is a minimum of 1 and a maximum of 3 relevant tags per recording. The Recall@K is estimated for each recording and averaged across all recordings in the dataset. Formally,

$$Recall@K = \frac{1}{N} \sum_{i=1}^{N} \frac{r_i}{k_i} \tag{7.1}$$

where N is the total number of recordings, r_i is the number of relevant tags returned in the top $K = 10$ predictions of recording i, and k_i is the number of true tags for the recording bounded by 1 and 3 as explained above.

The accuracy of music tagging with the above two CNN models is compared with the accuracy of MFCC features. MFCC-based systems are a common baseline approach for CNNs trained from Mel spectrograms where the desired CNN has to learn more than just the MFCC derivation. This is also the baseline approach considered in related research (Choi et al., 2016).

Another approach considered for the evaluation of the CNNs is the comparison with the Linear Discriminant Analysis (LDA) features described in Chapter 6. In Chapter 6 features of rhythm, timbre, melody, and harmony

extracted from Mel spectrograms and chromagrams and transformed with LDA achieved optimal performance in a country classification task (Section 6.3.2). The music tagging systems described above learn music similarity relations from the Mel spectrograms and enhanced metadata (language and decade combined with country information). This additional evaluation aims to compare the CNN with the LDA approach. For this evaluation, the dataset of 8200 recordings from Chapter 6 is used, and LDA features are compared to CNN features (taken from the penultimate layer of the networks, Figure 7.1). In particular, adding LDA melodic and harmonic descriptors extracted from chromagrams can be considered complementary to the CNN features derived from Mel spectrograms. The LDA classifier as found optimal in Section 6.3.2 is used. The train and test sets for the classification task are derived from the dataset of 8200 recordings as described in Section 6.2.4 but the test set is modified to remove any instances that overlap with the train set of the music tagging system (a subset of the dataset with 18054 recordings). Although this modification results in a slightly smaller test set (see also Section 7.2.1), the classification accuracy for the LDA features as reported in Section 7.3 is almost the same as the accuracy reported in Section 6.3.2. The random baseline predicting 1 of 134 countries given the distribution of countries in the modified test set is estimated by $\sum_{c \in C} I_c^2$ for I_c the proportion of recordings from country c compared to the total number of recordings, and C the set of all countries.

The combination of LDA with CNN features is also tested since the LDA features include attributes derived from chromagrams (e.g., the melodic and harmonic descriptors, Section 6.2.3) which can be considered complementary to CNN features derived from Mel spectrograms. In particular, CNN features from the best performing model (CNN-4L) are tested with combinations of: a) LDA melodic features (denoted CNN-4L+LDA-mel for the CNN-4L model), b) LDA harmonic features (denoted CNN-4L+LDA-har), c) LDA rhythmic features (denoted CNN-4L+LDA-rhy), d) LDA timbral features (denoted CNN-4L+LDA-tim), e) LDA (all) features (denoted CNN-4L+LDA-all) and combinations of the above.

7.2.4 Modelling expectations

Music similarity expectations are derived from the metadata of recordings and focus on geographical and cultural links. In Chapter 6 the assumption that recordings from the same country exhibit a similar music style was considered. Similar to this assumption, recordings from neighbouring countries might also exhibit similarities in their music. This is an assumption partly made in the classification of music by broader geographical areas (Gómez et al., 2009; Kruspe

et al., 2011). In this chapter, recordings from the same country neighbourhood as derived with the approach described in Section 6.2.6, and summarised in Table A.1, are expected to be similar.

Language is another cultural identifier and it has been used to distinguish music styles between different populations (Brown et al., 2014; Le Bomin et al., 2016). In this case, recordings exhibiting the same language could also be expected to have similar music content. Modelling language could be particularly useful for studying migration patterns. For example, music recordings from Canada include English as well as French examples and one might expect that the French examples are more similar to music from France whereas the English examples are more similar to music from the United Kingdom.

Decade information is included in the training set of the music similarity model (Section 7.2.1), where it is assumed that the combination of spatial and temporal information enhances the notion of music style and the derivation of music similarity relations. For example, music from a given country recorded in the 1950s might exhibit a different style than music from the same country recorded 40 years later. However, it is not expected that two recordings sharing only their decade, but no other geographical or cultural information, would be similar. Therefore, while decade is included in the dataset of the music tagging system it is excluded from the ground truth of music similarity expectations.

Overall, two recordings are expected to be similar based on their metadata if: a) the recordings come from the same country, or b) come from neighbouring countries (for the spatial neighbourhoods denoted in Table A.1), or c) have the same language. These are the music similarity expectations taken into account before concluding that two recordings are unexpectedly similar.

7.2.5 Unexpected similarity

Music similarity expectations derived from the metadata are compared with audio content similarities estimated from the CNN features. In particular, the penultimate layer of the best performing CNN model, as estimated via the evaluation described in Section 7.2.3, outputs the audio features for content similarity. The CNN features are aggregated with LDA features as found optimal in the results below (Table 7.3). Feature vectors are compared with cosine distance and for each recording the top K neighbours are kept for further analysis.

The dataset for the analysis of unexpected similarities is the set of 8200 recordings as defined in Chapter 6. This dataset is preferred over the dataset used to train the music tagging system (Section 7.2.1) because it provides a wider geographical spread of world music (see also the justification in Section 7.2.1). Considering the audio content-based distance and the music similarity expecta-

tions defined above, two recordings are considered 'unexpectedly similar' if: a) their audio feature vectors are close in the space according to the cosine distance, and b) they don't share any of the (expected) spatial and cultural metadata as described in Section 7.2.4.

The nearest $K = 5$ neighbours are selected for further analysis. In an ideal scenario it is expected that the most similar recordings will originate from the same country, so evaluating only the nearest $K = 1$ neighbour might always give the 'expected similarity'. In order to look for unexpected similarities, more nearest neighbours need to be evaluated. Selecting too many neighbours however will also result in irrelevant similarities because the distances between recordings may become too large too quickly as is often the case in high-dimensional spaces (Aggarwal, 2005). The choice of $K = 5$ is made because it evaluates enough neighbours to allow for unexpected similarity patterns to be revealed and it bounds the amount of neighbours by a number smaller than the minimum number of recordings per country (minimum 10 recordings per country as stated in the dataset description, Section 6.2.1).

Evaluating similarity in the dataset of 8200 recordings with 5 nearest neighbours for each recording results in a total of 41000 pairs. Each pair of recordings is annotated as unexpectedly similar, if the recordings share no metadata as explained above, and expectedly similar otherwise. The total number of unexpected pairs is compared to the random baseline and additional analysis focuses on summarising the unexpected similarities between countries. For example, for recordings of a given country, unexpectedly similar pairs are computed and the distribution of countries of the unexpectedly similar recordings is summarised.

In a trivial scenario, when the audio content of recordings from a given country is homogeneous and no unexpected similarities are found, each recording is similar to 5 other recordings from the same country. In a random scenario, i.e., in the case where audio content similarity is random, the distribution of unexpectedly similar countries approximates the frequency of each country in the dataset. While some randomness is expected considering the diversity of music styles within a country as well as hubness effects from the high-dimensional space (Schnitzer et al., 2012), consistent peaks in the distribution of unexpectedly similar countries might uncover hidden country relations. The pairs of countries with the most unexpected similarities are investigated.

Model	AUC	Recall@K
CNN-2L	0.881	0.606
CNN-4L	**0.885**	**0.608**
MFCC	0.865	0.584

Table 7.2: Evaluation for the different model architectures presented in Section 7.2.2 based on the AUC and Recall@K metrics as described in Section 7.2.3.

7.3 Results

7.3.1 CNN validation results

Tag predictions from the different CNN architectures (Table 7.1) are evaluated with the AUC and Recall@K metrics as described in Section 7.2.3. Results are compared to the baseline MFCC system as shown in Table 7.2. The four layer architecture (CNN-4L) achieved best performance considering both these metrics.

The prediction accuracy for each tag for the best performing CNN model (CNN-4L) was also assessed with the AUC metric as shown in Figure 7.4. The highest accuracy was predicted for the decade tag 1910 (AUC=0.998) whereas the decade 1940 was the $10th$ most accurate tag. The $2nd$ to $9th$ most accurate tags included languages and countries from mainly African regions. These tags were, in order of most to least accurate, the Ga language of Ghana (AUC=0.998), the country tags Zambia (AUC=0.996), and Saint Lucia (0.994), the Acholi language of Uganda (AUC=0.986), countries Iraq (AUC=0.981), Sudan (AUC=0.981), Nepali language (AUC=0.963), and South Sudan (AUC=0.963). Amongst the least accurate tags with AUC score near random (AUC$\approx$ 0.5) are the countries of Haiti, Angola, Tanzania, Romania, and Senegal, the languages Southwestern Dinka and Spanish, and the decades 2000, 1950, and 1970. Low accuracies are expected when different music styles share the same tag. For example, decade tags are not expected to be indicative of the music style since many recordings from all over the world may have been recorded in the same decade. An exception holds for the decades 1910 and 1940 in this dataset because only a small number of recordings from limited geographical regions corresponded to these decade tags hence they also scored a high AUC. In particular, recordings with the 1910 decade tag originate from Nigeria, India, and Pakistan and they comprise audio material digitised from the Ethnographic Wax Cylinders collection. Due to the nature of this collection a characteristic noise is prominent in the background of these recordings creating an "era effect" that can be easily captured by the CNN model.

The models were also evaluated for the task of country prediction using the

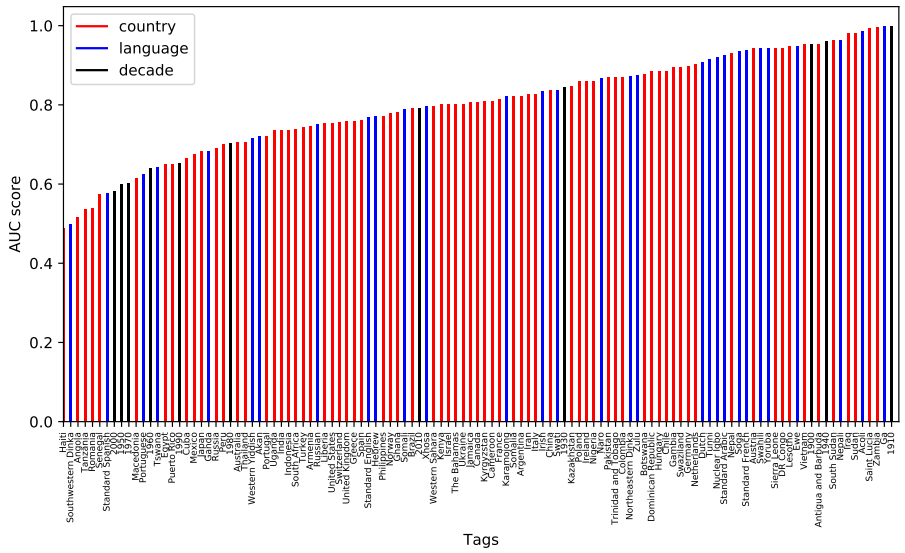


Figure 7.4: The prediction accuracy per tag estimated via the AUC metric for the best performing model (CNN-4L, as shown in the results of Table 7.2).

penultimate CNN layer as audio features as described in Section 7.2.3. The dataset of 8200 recordings from Chapter 6 is used and the performance of the CNN models is compared with the previously reported performance of the LDA features (Section 6.3.2) as well as combinations of the CNN with the LDA features. The test set for the classification task is slightly modified as described in Section 7.2.3. The random baseline predicting 1 of 134 countries given the distribution of countries in the modified test set is 0.010. Results from the country classification task are shown in Table 7.3.

As shown in Table 7.3 the CNN-4L model performs better than the CNN-2L (F-score of 0.122 compared to 0.083) but still poorer than the LDA features (F-score of 0.325). When LDA features are concatenated with CNN features the classification accuracy is improved. In particular adding LDA rhythmic, timbral, and harmonic features to the CNN-4L features increases the accuracy of the CNN-4L by $\approx 20\%$. The highest accuracy (0.337) is achieved when all LDA features are appended to the CNN-4L features. The low accuracy of the CNN models compared to the LDA model could be attributed to the fact that the CNN models were trained with only a subset of the country labels found in the test set. In particular, the CNN models were trained on a set of recordings from 80 countries (and additional language and decade tags as explained in Section 7.2.1) whereas the LDA model was trained on a set of recordings from 137 countries (Section 6.2.1). Note that many of the countries included in the test set (with a total of 134 countries) were not present in the training set of

Model	F-score
CNN-2L	0.083
CNN-4L	**0.122**
LDA-all	0.325
CNN-4L+LDA-mel	0.168
CNN-4L+LDA-har	0.178
CNN-4L+LDA-rhy	0.215
CNN-4L+LDA-tim	0.202
CNN-4L+LDA-mel-har	0.198
CNN-4L+LDA-mel-har-rhy	0.291
CNN-4L+LDA-rhy-tim	0.301
CNN-4L+LDA-rhy-tim-har	0.332
CNN-4L+LDA-all	**0.337**
Random baseline	0.010

Table 7.3: Class-weighted accuracy for the classification of recordings by country combining CNN features as explained in Section 7.2.3 and LDA features as explained in Section 6.2.4. The LDA classifier is used since it outperformed other classifiers (Section 6.3.2).

the CNN model but they were present in the training set of the LDA model. The differences in the performance between the CNN and LDA models could be evaluated further in future work.

7.3.2 Unexpected similarity findings

The best performing model, in this case the CNN-4L combined with the LDA features of rhythm, melody, timbre, and harmony, is used to study unexpected similarities. Pairwise similarities between audio feature vectors of recordings are estimated with the cosine distance. The nearest 5 neighbours of each recording are considered and their metadata are compared. From a total of 41000 pairs (8200 recordings with 5 neighbours each) a total of 28626 pairs are found to be similar in their audio content but originate from different countries. When language information is considered in addition to country, the number of unexpectedly similar pairs of recordings reduces to 28323. Language information has a small effect in the number of unexpectedly similar pairs compared to country information (28626 reduced to only 28323) but this is because language tags are highly correlated with country tags. Adding spatial neighbourhoods (Section 6.2.6 and Table A.1) in the similarity expectations reduces even further the number of unexpectedly similar pairs. In particular, comparing whether similar recordings originate from the same or neighbouring country gives a total of 26672 unexpectedly similar pairs, and adding language information to the

Figure 7.5: Unexpected similarities between countries (only the most consistent similarities between pairs of countries are shown).

expectations of country and spatial neighbourhoods results in a total of 26504 unexpectedly similar pairs.

The set of 26504 unexpectedly similar pairs after evaluating similarity in the country, spatial neighbourhood, and language metadata of each recording, is considered for further analysis. In particular, the unexpected similarities between countries are investigated. Figure 7.5 shows the links between countries with the most unexpectedly similar recordings. In particular a line is drawn between recordings of country A that were found (unexpectedly) similar to recordings of country B. Even though recordings from country A might be similar to recordings of many other countries, only the country of maximum unexpected similarity (in this example country B) is selected for visualisation purposes. What is more, pairs of unexpected countries are plotted when the similarity is consistent, i.e., it occurs in more than a percentage r of comparisons of recordings from country A. The percentage r of comparisons is empirically set to 9 such that only the most consistent pairs of countries are selected for further analysis. For a country with 10 recordings (the minimum) and 5 nearest neighbours for each recording, the 9% threshold amounts to 4 unexpectedly similar pairs out of 50. For a country with 100 recordings (the maximum), it amounts to 45 out of 500.

Based on the results shown in Figure 7.5 the following can be summarised. Recordings from Iceland are unexpectedly similar to recordings from Bulgaria (12.8% of unexpected pairs from the total number of recording pairs compared for Iceland), Costa Rica to Antigua and Barbuda (12.7%), Suriname to Somalia (11.6%), Republic of Serbia to Bulgaria (11.3%), Denmark to Poland (10.0%), Saudi Arabia to Swaziland (10.0%), Paraguay to Germany (9.6%), Antigua and Barbuda to Trinidad and Tobago (9.5%), French Polynesia to Philippines (9.3%), South Korea to Sudan (9.1%), and Chad to South Africa (9.1%) (in

order of most to least unexpected pair rate per country).

Listening to music examples the following timbral characteristics can be summarised for the unexpectedly similar pairs. The recordings from Iceland found similar to recordings from Bulgaria are choir singing examples (Bulgarian recordings are liturgical chants) in slow tempo with reverb and male and female singing voices with wide pitch range. Recordings from Costa Rica found similar to recordings from Antigua and Barbuda exhibit similar harmonies linking guitar examples from the former country with steel drum (a chromatically pitched percussion instrument) examples from the latter. Recordings from Suriname feature the music of Saramaka people of African descent and they were found similar to music from Somalia. These recordings featured male voices alternating between slow and fast syllabic singing (the latter sounded similar to speech). Recordings from the Republic of Serbia similar to recordings from Bulgaria featured choir singing examples (similar to the characteristics of the Iceland-Bulgaria pairs described above) linking chorus music from the era of Yugoslavia to liturgical chants from Bulgaria (with origins also in Russia and Ukraine). Recordings from Denmark similar to recordings from Poland featured accordion music as well as examples of singing accompanied by guitar, whereas recordings from Saudi Arabia similar to recordings from Swaziland featured group singing in slow tempo and low pitch. Other unexpected similarities linked Central African to South African regions, for example Chad to South Africa (with prominent drums accompanying vocals), as well as islands in relative geographical proximity, for example French Polynesia to Philippines in the Pacific (with polyphonic orchestra accompanying vocals), and Antigua and Barbuda to Trinidad and Tobago in the Caribbean (with steel drum examples). Listening examples can be accessed via the interactive visualisation of Figure 7.5 available online at `http://mpanteli.github.io/phd_thesis/chapter_7/unexpected_similarity.html`.

7.4 Discussion

In this chapter, a music tagging system was developed with convolutional neural networks. The system was trained to predict tags of countries, languages, and decades from the Mel spectrograms of recordings. Following common approaches in state-of-the-art research two CNN architectures were tested. The tag prediction accuracy achieved results higher than the baseline MFCC system and an overall AUC accuracy of 0.885 that is comparable to the performance of Western music tagging systems (for example the highest AUC=0.894 for the Western music tagging system developed by Choi et al. (2016)).

The performance of the model was also assessed with a classification task

as in Section 6.3.2 predicting the country of a recording for the dataset of 8200 recordings (Section 6.2.1). While the CNN features on their own did not perform better than the LDA features for the country classification task, the combination of CNN with LDA features increased the previous state-of-the-art accuracy (achieved by the LDA features alone) by approximately 1%.

The study of unexpectedly similar music revealed some interesting links between Icelandic, Bulgarian, and Serbian music, especially regarding their choir singing examples. It also revealed similar music examples between the countries of Suriname (for music of Saramaka people) and Somalia, two very distinct geographical regions but with populations of perhaps similar origin (the Saramaka people of Suriname are of African descent). Other unexpected similarities linked music between islands in the Pacific and between islands in the Caribbean.

This is the first music tagging system designed specifically for world music. The pre-trained model can be used in the future to label countries, languages, and decades of world music recordings with incomplete metadata. The CNN features can be used to model music similarity in other datasets in the same manner as described in this chapter.

Many aspects of the methodology can be improved in future work. Mel spectrograms were used as input to the CNN but chromagrams might be a more suitable input to capture pitch attributes (Bittner et al., 2017). Evaluating similarities from the CNN with LDA features indicated a relatively high weight on timbral characteristics (see also the music examples discussed in Section 7.3). Adding harmonic and melodic features from a CNN model trained on chromagrams might reveal similarities beyond prominent timbre characteristics.

Only two CNN architectures were investigated in this chapter but different configurations could be further explored. CNNs were selected following state-of-the-art music tagging research but other deep learning models such as recurrent neural networks (RNNs) can be considered. RNNs are more suitable for modelling the temporal relationships of the signal compared to the spatial properties of the spectrogram captured by the CNNs (Humphrey et al., 2013). Alternatively, the combination of both CNNs and RNNs, can be considered (Choi et al., 2017b). The target labels in this work were treated as high-dimensional sparse binary vectors but low-dimensional embeddings can also be derived, for example with matrix factorisation techniques as proposed in related research (Van Den Oord et al., 2014; Oramas et al., 2017).

The number of unexpectedly similar pairs was relatively high (around 64% of the total number of pairs was identified as unexpectedly similar). This can be due to the lack of metadata. For example, country, language, and spatial neighbours were used as ground truth for similarity but additional metadata including the culture or ethnic group or the genre of a song could be considered

in this comparison. What is more, language tags were only present in 28% of recordings and this lack of information could increase the probability of falsely identifying two recordings as unexpectedly similar. Effects of hubness might also be the cause of detecting too many unexpectedly similar pairs. For example, high-dimensional data have the tendency to contain points that frequently occur as nearest neighbours of other points. Therefore the 5 nearest neighbours for each recording might not be a result of true content similarity but rather artefacts from high dimensionality.

Similarity was assessed for $K = 5$ nearest neighbours for each recording. The optimal number of K can be investigated further taking into account hubness effects. The unexpectedly similar pairs presented above indicated cultural and geographical links between countries but more can be explored to explain the musicological or historical reasons behind these links.

7.5 Outlook

In this chapter a model for music similarity was developed extracting audio features in a data-driven approach. In particular, a convolutional neural network was trained to learn mappings between Mel spectrograms and music tags such as the country, language or decade of the recording. The penultimate layer of the network was used as the feature vector and cosine distance was applied to estimate the audio content similarity between recordings. Similarity expectations were modelled from spatial and cultural metadata. Recordings that appeared to have similar audio content (as estimated via the cosine distance of audio embeddings) but shared no ground truth (as defined by their metadata) were termed as unexpectedly similar. By tracking the origins of these unexpectedly similar pairs, observations were made regarding possible links between the music of different countries. The musicological interpretation of these results can be investigated further in future work.

This chapter focused on CNNs, a state-of-the-art approach in music content description today (Choi et al., 2016; Schlüter, 2016; Bittner et al., 2017). While music tag predictions achieved overall high accuracies, the mapping from Mel spectrograms to musically relevant dimensions could be examined further. For example, the model can be tested for overfitting and sampling bias and the non-linear projections could be evaluated with respect to desired music properties (Sturm, 2013; Mishra et al., 2017).

This chapter complemented the developments of the previous chapters by placing the focus on music similarity and assessing data-driven methods that advance the custom feature design process. Several steps in the proposed methodologies could be improved and further analyses could expand the findings of

this thesis. The following chapter discusses directions for improvement in future work and presents the concluding remarks.

Chapter 8

Future work and conclusion

In this thesis, world music corpora were analysed with computational tools and research questions on music similarity and dissimilarity were addressed. The methodology relied on audio features extracted automatically from sound recordings and data mining techniques applied to quantify similarity relations. Considerations were made on the suitability of audio features for the study of world music similarity and an evaluation strategy was proposed to assess the robustness of the features to particularities of world music recordings. Hand-designed features were used to model singing voice attributes and clustering techniques assessed the similarity of singing styles across music cultures. Audio features were extracted with a combination of hand-designed and machine learning methods and outlier detection was used to assess relations of music dissimilarity. Lastly, a data-driven approach with convolutional neural networks was used to derive audio features from Mel spectrograms and contrasting similarities in the audio and metadata domains revealed unexpected links between world music cultures.

The proposed methodologies were specifically designed to enable computational analysis at large scale while at the same time limiting the Western music bias (Section 4.3). The data augmentation technique discussed in Section 4.4 focused on challenges found in world music corpora and contributed to the evaluation of existing audio features for the study of similarity in world music. The audio features used in Chapter 5 were specifically designed for the task of pitch content description whereas the audio features used in Chapters 6 and 7 were automatically learned from low-level representations of the world music recordings. These approaches are more suitable for world music content description and compared to other methods proposed for the analysis of world music (Lomax, 1976; Savage et al., 2012), they can scale with even larger datasets.

The findings discussed throughout the thesis are subject to the particulari-

ties of the analysed datasets. The amount of outliers detected per country was not significantly correlated with the amount of recordings from that country (Section 6.5) and music tags predicted with high accuracies did not correspond to the most popular tags in the dataset (Section 7.3). However, pairwise similarities between recordings as described in Sections 5.3.3 and 7.2.5 can be affected by slight changes in the dataset, especially for music cultures represented by a small number of recordings. More work needs to be done on establishing the extent to which the findings in the aforementioned experiments generalise to other world music datasets. Directions for future improvement are discussed in more detail in the sections below.

8.1 Limitations of the dataset

The developments in this thesis and the musicological insights summarised in Chapters 5, 6, and 7 rely on a subset of the world music corpus (Section 3.1). While attempts have been made to select a geographically diverse and large corpus the question of how representative the corpus is still remains unanswered. In particular, drawing randomly a maximum of 100 recordings per country, as used in the dataset in Chapter 6 for studying music outliers, reduces the possible sampling bias compared to, for example, manually selecting these recordings, but doesn't guarantee that the selected recordings are the most representative of the music of a country. In fact, the archives themselves don't provide any guarantee that their collections are representative of world music and the metadata associated with each collection lack information that could be used to assess the quality of each sample.

The world music recordings from the Smithsonian Folkways Recordings and the British Library Sound Archive are compiled from individual collections of musicological or other research. The primary purpose of each collection is not necessarily the comparison of music cultures as considered in this thesis. Therefore the available recorded material might not be suitable for addressing comparative music research questions. What is more, the criteria followed for the music curation in the above two archives are not explicitly stated. Looking at the geographical representation of the music of each archive some bias can be observed. For example, the World and Traditional Music Collection of the British Library Sound Archive consists of a large representation from the United Kingdom and other British ex-colonial regions such as India and South Africa. The world music collection from the Smithsonian Folkways Recordings consists of a large representation of recordings from the United States including music from its minority populations such as Jewish, Italian, and Indian as well as music from its neighbouring countries (Canada and Mexico) and South America.

Assessing how representative a collection of music is of the music in a particular region or culture is a challenging task. Expert knowledge of the music is required in order to include all different music styles within a region and a large set of recorded material covering all music needs to be available before selecting the final sample. This would require new recorded material to be created or collected, especially for countries or regions where only a small set of recordings was available and this was from a very specific area or music style that doesn't generalise well to the music of the whole region. For the purpose of a music comparative study, criteria based on music diversity and completeness could be addressed. For example, the recordings collected from a specific region need to capture all different styles of that region and together need to form a complete set with a fair distribution over the different styles. Defining what constitutes a 'complete' set from a comparative perspective is difficult but consulting music experts from each region or culture could give a better overview of the individual music collections that together constitute the world music corpus.

Additional metadata could also be useful to assess how representative each sample is. For example, the exact location of the recorded material or the music style (if known) can be considered in the metadata for a balanced representation of the music from a given region. What is more, the primary purpose of the collector recording such material could be included in the annotations as a point of reference when creating a world music corpus for the comparison of music cultures. Ideally, each smaller collection of recordings is put together with a comparative music purpose in mind or with criteria that could generalise to this comparative music application. If instead a collection of recordings was created to study a very specific music style from a given region then these recordings would not be ideal for music comparative purposes as they are not diverse nor complete for the music of that region. For folk music, descriptions such as the social function of a song, i.e., being a love, dance or war song Shanahan et al. (2016), can be used to balance the representation of different styles of folk songs.

All the above limitations need to be taken into account when looking at the results in Chapters 5, 6, and 7. For example, in the dataset used in Chapter 6 some countries were only represented by 10 recordings, with all 10 recordings possibly originating from the same village or group of performers. This is not enough to capture the diversity of music from this country and possible biases could be observed in the detection of outlier recordings. In addition, unexpectedly similar pairs observed in Chapter 7 might be due to peculiarities of a few recordings in the dataset rather than actual similarities between the musical cultures considered.

The aforementioned limitations of the dataset affect the development of music similarity models and their applications as described in Chapters 5, 6, and 7.

Future work to improve this kind of data and research framework is described below.

8.2 Future work

The first consideration for future work is the expansion and improvement of the world music corpus (Section 3.1). The study of music dissimilarity (Chapter 6) and unexpected similarity (Chapter 7) focused on a dataset of 8200 recordings from 137 countries derived from the BLSF corpus (Section 3.1). Although this is the largest world music dataset studied so far, several geographical regions and cultures were not represented. This dataset, reflecting the geographical distribution of BLSF, lacked examples from the majority of African countries such as Mauritania, Libya, Niger, Central African Republic, and Namibia. It also included an over-representation of Western and Western-colonial regions. The corpus can be expanded by adding ethnomusicological collections from other music archives. For example, the music archive of the Centre de Recherche en Ethno-Musicologie (CREM)[1] in France provides a good representation of world music cultures with more than 7000 hours of audio data (Fourer et al., 2014) that could complement the geographical representation of the current world music corpus. Music from Africa could also be expanded with recordings from the audio archive of the Royal Museum for Central Africa (RMCA)[2] in Belgium which consists of more than 3000 hours of audio recordings from Central Africa (Moelants et al., 2009). The creation of a representative world music corpus will continue indefinitely as more music is recorded and the digitisation of archived recordings proceeds.

The derived world music corpora (Section 3.2) for each study (Chapters 5, 6, 7) were created by sampling recordings based on (mainly) country information. That is, the corpus criteria focused on balancing the representation of recordings from each country, where country was considered a proxy for music style. In Chapter 7 additional metadata on the language and year of the recording (whenever available) were considered for an improved approximation of music similarity. The metadata associated with each collection can be improved and models could be trained to learn from a variety of music style indicators. For example, ethnic group and language information could be good identifiers for music culture (Brown et al., 2014; Le Bomin et al., 2016) and countries could be further divided into sub-regions for a more fine-grained approximation of world music styles.

Metadata like the recording equipment, recording date, performer or collec-

[1] `http://archives.crem-cnrs.fr/`, accessed 15th November 2017.
[2] `http://music.africamuseum.be`, accessed 15th November 2017.

tor could also be used to model audio quality effects in the estimation of music similarity. In particular, low-level descriptors might be capturing attributes of the recording quality in addition to the desired music signal properties (Sturm, 2013; Urbano et al., 2014). What is more, if a country or music style is represented in majority by the same performers then the music system might be learning specific aspects of the performer's style that do not generalise to the overall music style (Flexer and Schnitzer, 2010). In this thesis, effects of recording quality were assessed for the rhythmic and melodic descriptors used in the music dissimilarity model (Section 4.4 and Chapter 6). The development of a singing style similarity model (Chapter 5) was based on mid-level descriptors that are quite robust to these effects (Flexer and Schnitzer, 2010). In future work, audio quality effects could be modelled for the music similarity model presented in Chapter 7 as well as the timbral and harmonic descriptors used in the music dissimilarity model (Chapter 6). Expanding the corpus and balancing the performer distribution in the representation of each music culture would also limit the album-artist effects in future work.

There is a semantic gap between manual music annotation as proposed in ethnomusicological research (Savage and Brown, 2013) and computational descriptors as proposed in MIR research (Gómez et al., 2013). In Chapters 2 and 4, common descriptors and terms from both ethnomusicology and MIR were reviewed. For the study of world music styles in particular, ethnomusicologists have developed specific annotation systems (Lomax, 1976; Savage et al., 2012) to capture particularities between different world music cultures. Such an annotation system does not exist for MIR research, although some algorithms have been proposed for the analysis of specific non-Western music cultures (Serrà et al., 2011; Åđentürk et al., 2013; Srinivasamurthy et al., 2014; Ganguli et al., 2016). The majority of computational studies for the comparison of world music corpora (including the approaches presented in this thesis), have focused on using existing MIR tools (built with the primary aim of Western music analysis) and slightly adaptating them for non-Western music (Gómez et al., 2009; Kruspe et al., 2011; Panteli and Dixon, 2016). A computational toolbox that models the descriptors defined in related ethnomusicological research such as the Cantometrics or Cantocore systems (Lomax, 1976; Savage et al., 2012), would be beneficial. This would aid ethnomusicological research by replacing the time-consuming, and often subjective, task of manual annotation with efficient computational processing. The challenge in this case would be to design computational descriptors that accurately represent the high-level properties defined by ethnomusicologists (Huron, 2013; Wallmark, 2013) and are in agreement with music perception and cognition models (Desain et al., 1998).

All the approaches proposed in this thesis for audio feature extraction ig-

nored at a large part the overall temporal structure of the signal. In Chapter 5 pitch descriptors were summarised across the whole duration of the recording, in Chapter 6 music elements were modelled for short overlapping windows of 8 seconds, and in Chapter 7 2-dimensional convolutions modelled music aspects over relatively small frequency and time frames. Several approaches have been proposed in the literature to explicitly model the temporal relations of a music recording, including for example, Hidden Markov Models (Ni et al., 2012), Viterbi decoding (Klapuri et al., 2006), and neural network approaches such as Long Short Term Memory models (Humphrey et al., 2013) and Recurrent Neural Networks (McFee and Bello, 2017). These could be further explored and appended to the methodology proposed in this thesis.

Musicological findings presented in this thesis were evaluated qualitatively via listening to music examples. Additional quantitative evaluation could be performed to validate the findings in Chapters 5, 6, and 7. In particular, the singing clusters in Chapter 5 could be evaluated further by studying which pitch descriptors are prominent in each cluster. Musical features prominent in the outlier recordings in Chapter 6 could also be quantitatively compared and validated. For example, the use of polyrhythms was observed as a significant feature for rhythmic outliers. A descriptor could be designed to annotate the use of polyrhythms and the distribution of this descriptor across recordings identified as rhythmic outliers could be compared to the non-outlier recordings. Likewise, recordings from China, the country with the most spatial outliers, were observed to consist of bright timbres. A descriptor capturing the energy in high frequency bands, for example the spectral centroid, could be computed for recordings of China detected as outliers and compared to recordings from its neighbouring countries. This would validate the statement that bright timbre is a prominent feature in the outlier recordings from China.

This work focused on musicological research questions regarding similarity relations in world music and proposed a variety of computational methods to address them. While computational methods are able to analyse large corpora, the lack of expert knowledge in the derived music descriptors could be considered a disadvantage (Fink, 2013). Qualitative evaluation was partly considered to account for the absence of a music similarity ground truth. In particular, outlier detection was assessed with a listening experiment in the odd one out fashion (Section 6.4) and music characteristics were summarised in each study after listening to the music examples (Sections 5.3.3, 6.3.3, 7.3). Observations of musical exchange in the history of world music need to be verified and interpreted by experts in related fields including ethnomusicologists, anthropologists, or historians. Multi-disciplinary research could significantly advance the study of world music cultures and findings proposed in this thesis could form the basis

for future explorations.

8.3 Conclusion

The computational analysis of world music corpora has been considered in this thesis. Challenges of processing a corpus of world music recordings were discussed and three different methodologies were proposed to explore similarities in world music. This thesis advanced the state-of-the-art of MIR methods for the analysis of world music recordings and contributed musicological insights with respect to the way music cultures are different and alike.

One of the contributions of this thesis is the curation of the largest world music corpus for computational analysis. More specifically, recording collections from the Smithsonian Folkways Recordings and British Library Sound Archive were combined to create the BLSF corpus of world music data for approximately 60000 recordings (Section 3.1). The associated metadata were curated to remove inconsistencies and improve partly labelled annotations (Section 3.1.1). The sound recordings were processed in the form of Mel spectrograms (Section 4.1.1) and are made available to other researchers for the continuation of this line of research. The study of this world music corpus with computational tools enabled musicological discoveries at a larger scale than had been previously addressed.

While many automatic systems have been developed for the analysis of Western and non-Western music, little has been done to assess how well existing algorithms perform on a dataset of world music recordings. An evaluation strategy was proposed to assess the suitability of rhythmic and melodic descriptors for the study of similarity in world music corpora (Section 4.4). The evaluation strategy shared concepts with other MIR studies on data augmentation but focused on creating synthetic audio data that reflect the challenges found in world music corpora. The derived dataset is available for the continuation of research on evaluating MIR tools for world music processing.

The thesis advanced the state-of-the-art of MIR algorithms for world music analysis. In particular, developments targeted three different methods. First, the custom feature design of pitch descriptors was shown to be successful in capturing aspects of singing style similarity in world music. Second, the combination of low-level music descriptors with linear dimensionality reduction techniques efficiently adapted audio representations to a musically relevant space. This was an improved method from the previous study on singing as expert knowledge was only partly required and higher-level dimensions were learned directly from the world music data. Lastly, a convolutional neural network was trained directly from the Mel spectrograms of sound recordings to predict world music tags. This method explored non-linear projections of the data and limited

expert knowledge was required during training. The pre-trained model is one of the first automatic world music tagging systems and can now be used to assist in tagging of partly unlabelled world music collections.

A variety of data mining tools were considered for measuring aspects of similarity and dissimilarity in world music. While only existing data mining methodologies were used, the combination of these tools with the audio feature extraction proposed a new way for the analysis of world music. More specifically, unsupervised clustering aided the discovery and exploration of singing styles in world music. Outlier detection techniques formed the basis for quantifying music dissimilarity and distance metrics in complex manifolds revealed structures of music similarity.

The computational approaches allowed for music comparisons at a large scale uncovering patterns of music similarity across the world. More specifically, singing style clusters revealed characteristic uses of vibrato and melisma and the use of slow versus fast syllabic singing (Chapter 5). Eastern Mediterranean cultures were found to have similar singing styles with prominent use of melisma whereas central and northern European cultures were grouped together with prominent use of vibrato. Both cultures differed from African and Caribbean singing styles characterised by fast syllabic singing.

The study of music dissimilarity revealed geographical regions that have distinctive music examples (Chapter 6). The country of Botswana had the largest number of music outliers compared to the whole dataset and these outliers exhibited distinctive uses of rhythms and harmonies. China had the most music outliers compared to its neighbouring countries and these outliers featured bright timbres and singing in high frequencies. With respect to rhythmic outliers, African countries such as Benin and Botswana had the most outliers with recordings, featuring the use of polyrhythms. For harmonic outliers, south East Asian countries such as Pakistan and Indonesia, and African countries such as Benin and Gambia, were found to have the most outliers with recordings featuring the use of inharmonic instruments such as gongs and bells.

The study of music similarity (Chapter 7) used audio features derived from a convolutional neural network combined with the previously extracted Linear Discriminant Analysis features (Section 6.2.3) and revealed similarities linking the music of Saramaka people (of African descent) of Suriname with Somalia, instrumental music from Costa Rica with Antigua and Barbuda, and choir singing music from Iceland with Bulgaria. These similarities could not be justified by spatial or linguistic proximities as derived from available metadata (Section 7.2.4).

The study of world music as proposed in this thesis is challenging because there is scattered information about the relationships between different music

cultures but no single source outlines all possible connections. Emphasis was placed on explicitly evaluating the computational tools, or otherwise demonstrating their efficiency, prior to applying them to the actual data. Western music bias may still have had an influence in the decisions made in the methodology and the interpretation of the results. Interdisciplinary perspectives could limit the risk of cultural bias and could open new directions to musicological research that could help us better understand musical exchange in the world.

While several comparisons have been addressed with computational tools, music has aspects that lie beyond the data representations used in this thesis. The developments in this thesis aim to assist musicological research but in no way replace the human contribution. As Clarke (2014, p. 12) puts it, "the empirical and the metric have as much potential as any other paradigm to work to humanistic ends, but the question is whether in anthropology, of all disciplines, the heart doesn't have a crucial role to play". The findings in this research can form a basis for future exploration and quantitative comparisons can be further expanded with qualitative studies for a more complete understanding of world music.

Appendices

Appendix A

Spatial neighbours

Country	Spatial neighbours
Afghanistan	Tajikistan, Pakistan, Uzbekistan, China, Iran
Algeria	Morocco, Tunisia, Mali, Western Sahara
Angola	Zambia, DR Congo
Antigua and Barbuda	Saint Lucia
Argentina	Uruguay, Chile, Bolivia, Paraguay, Brazil
Armenia	Azerbaijan, Turkey, Iran
Australia	Fiji, Papua New Guinea, Indonesia
Austria	Hungary, Germany, Czech Republic, Switzerland, Italy
Azerbaijan	Turkey, Russia, Armenia, Iran
Belgium	Germany, France, Netherlands
Belize	Guatemala, Mexico
Benin	Nigeria
Bhutan	India, China
Bolivia	Peru, Argentina, Chile, Paraguay, Brazil
Botswana	Zambia, Zimbabwe, South Africa
Brazil	Uruguay, Argentina, French Guiana, Bolivia, Guyana, Suriname, Colombia, Venezuela, Paraguay, Peru
Bulgaria	Romania, Macedonia, Greece, Turkey
Cambodia	Thailand, Laos, Vietnam
Cameroon	Gabon, Nigeria, Chad
Canada	United States of America
Chad	Nigeria, South Sudan, Cameroon
Chile	Argentina, Bolivia, Peru

China	Afghanistan, Kazakhstan, Kyrgyzstan, Laos, Russia, Bhutan, Mongolia, Myanmar, India, Tajikistan, Pakistan, Vietnam
Colombia	Panama, Ecuador, Venezuela, Peru, Brazil
Costa Rica	Panama, Nicaragua
Croatia	Hungary
Cuba	The Bahamas, Jamaica, Haiti
Czech Republic	Poland, Germany, Austria
DR Congo	United Republic of Tanzania, Angola, Rwanda, Zambia, South Sudan, Uganda
Denmark	Germany
Dominican Republic	Haiti
Ecuador	Peru, Colombia
Egypt	Israel, Sudan
El Salvador	Honduras, Guatemala
Ethiopia	South Sudan, Kenya, Sudan, Somalia
Fiji	Papua New Guinea, Solomon Islands, New Zealand
Finland	Norway, Sweden, Russia
France	Belgium, Germany, Italy, Switzerland, Spain
French Guiana	Suriname, Brazil
French Polynesia	Samoa, Mexico
Gabon	Cameroon
Gambia	Senegal
Germany	Poland, France, Austria, Belgium, Netherlands, Switzerland, Czech Republic, Denmark
Ghana	Ivory Coast
Greece	Bulgaria, Macedonia, Turkey
Grenada	Trinidad and Tobago, Antigua and Barbuda, Saint Lucia
Guatemala	El Salvador, Belize, Honduras, Mexico
Guinea	Liberia, Senegal, Sierra Leone, Ivory Coast, Mali
Guyana	Suriname, Venezuela, Brazil
Haiti	Dominican Republic
Honduras	El Salvador, Guatemala, Nicaragua
Hungary	Croatia, Romania, Ukraine, Austria
Iceland	Ireland, Netherlands, United Kingdom
India	Afghanistan, Bhutan, Myanmar, Nepal, China, Pakistan
Indonesia	Papua New Guinea, Malaysia
Iran	Afghanistan, Armenia, Azerbaijan, Iraq, Pakistan, Turkey
Iraq	Saudi Arabia, Jordan, Turkey, Iran

Ireland	United Kingdom
Israel	Egypt, Lebanon, Jordan
Italy	France, Switzerland, Austria
Ivory Coast	Liberia, Ghana, Mali, Guinea
Jamaica	Haiti, Cuba, The Bahamas
Japan	Philippines, South Korea
Jordan	Iraq, Saudi Arabia, Israel
Kazakhstan	Kyrgyzstan, Uzbekistan, Russia, China
Kenya	South Sudan, United Republic of Tanzania, Ethiopia, Somalia,Uganda
Kiribati	Guyana, Suriname, Brazil
Kyrgyzstan	Kazakhstan, Uzbekistan, China, Tajikistan
Laos	Thailand, Cambodia, Myanmar, Vietnam, China
Latvia	Lithuania, Russia
Lebanon	Israel
Lesotho	South Africa
Liberia	Sierra Leone, Ivory Coast, Guinea
Lithuania	Poland, Latvia, Russia
Macedonia	Bulgaria, Greece
Malawi	United Republic of Tanzania, Zambia, Mozambique
Malaysia	Thailand, Indonesia
Mali	Algeria, Senegal, Ivory Coast, Guinea
Malta	Italy, Tunisia, Greece
Mexico	United States of America, Belize, Guatemala
Mongolia	Russia, China
Morocco	Algeria, Western Sahara, Spain
Mozambique	United Republic of Tanzania, Zambia, Zimbabwe, Malawi, South Africa, Swaziland
Myanmar	Thailand, Laos, India, China
Nepal	India, China
Netherlands	Belgium, Germany
New Zealand	Fiji, Solomon Islands, Australia
Nicaragua	Costa Rica, Honduras
Nigeria	Cameroon, Benin, Chad
Norway	Finland, Sweden, Russia
Pakistan	Afghanistan, India, China, Iran
Panama	Costa Rica, Colombia
Papua New Guinea	Indonesia
Paraguay	Argentina, Bolivia, Brazil
Peru	Ecuador, Colombia, Chile, Bolivia, Brazil

Philippines	Malaysia, Vietnam, Indonesia
Poland	Lithuania, Germany, Czech Republic, Russia, Ukraine
Portugal	Spain
Puerto Rico	Antigua and Barbuda, Dominican Republic
Republic of Serbia	Croatia, Macedonia, Hungary
Romania	Bulgaria, Hungary, Ukraine
Russia	Kazakhstan, Poland, Finland, Latvia, Azerbaijan, Lithuania, Mongolia, China, Norway, Ukraine
Rwanda	United Republic of Tanzania, DR Congo, Uganda
Saint Lucia	Trinidad and Tobago, Antigua and Barbuda, Grenada
Samoa	Mexico, French Polynesia
Saudi Arabia	Iraq, Yemen, Jordan
Senegal	Gambia, Mali, Guinea
Sierra Leone	Liberia, Guinea
Solomon Islands	Fiji, Papua New Guinea
Somalia	Kenya, Ethiopia
South Africa	Zimbabwe, Mozambique, Swaziland, Botswana, Lesotho
South Korea	China, Japan, Philippines
South Sudan	Kenya, Ethiopia, DR Congo, Sudan, Uganda
Spain	Portugal, France, Morocco
Sudan	Egypt, South Sudan, Ethiopia, Chad
Suriname	Guyana, French Guiana, Brazil
Swaziland	Mozambique, South Africa
Sweden	Norway, Finland
Switzerland	Germany, Italy, France, Austria
Tajikistan	Afghanistan, Kyrgyzstan, Uzbekistan, China
Thailand	Malaysia, Cambodia, Myanmar, Laos
The Bahamas	Haiti, Cuba, Jamaica
Trinidad and Tobago	Venezuela, Grenada, Saint Lucia
Tunisia	Algeria
Turkey	Armenia, Azerbaijan, Greece, Bulgaria, Iran, Iraq
Uganda	South Sudan, Kenya, United Republic of Tanzania, DR Congo, Rwanda
Ukraine	Poland, Hungary, Romania, Russia
United Kingdom	Ireland
United Republic of Tanzania	Kenya, Rwanda, Zambia, Malawi, Mozambique, DR Congo, Uganda
United States of America	Canada, Mexico
Uruguay	Argentina, Brazil
Uzbekistan	Afghanistan, Kazakhstan, Kyrgyzstan, Tajikistan

Venezuela	Guyana, Colombia, Brazil
Vietnam	Cambodia, Laos, China
Western Sahara	Algeria, Morocco
Yemen	Saudi Arabia
Zambia	United Republic of Tanzania, Angola, Zimbabwe, Malawi, Botswana, Mozambique, DR Congo
Zimbabwe	Zambia, Mozambique, South Africa, Botswana

Table A.1: Spatial neighbours for each country in the dataset of 8200 world music recordings used in Chapter 6.

Bibliography

B. Aarden and D. Huron. Mapping European Folksong: Geographical Localization of Musical Features. *Computing in Musicology*, 12:169–183, 2001.

S. Abdallah, E. Benetos, N. Gold, S. Hargreaves, T. Weyde, and D. Wolff. The Digital Music Lab: A Big Data Infrastructure for Digital Musicology. *ACM Journal on Computing and Cultural Heritage*, 10(1), 2017. doi: 10.1145/2983918.

G. Adler. Umfang, Methode und Ziel der Musikwissenschaft [The Scope, Method and Aim of Musicology]. *Vierteljahresschrift für Musikwissenschaft*, 1(1):5–20, 1885.

C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, pages 901–909, 2005.

C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *Proceedings of the International Conference on Management of Data (ACM SIGMOD)*, pages 37–46, 2001.

J. J. Aucouturier, F. Pachet, and M. Sandler. "The way it sounds": Timbre models for analysis and retrieval of music signals. *IEEE Transactions on Multimedia*, 7(6):1028–1035, 2005.

J. Baily and D. M. Collyer. Introduction: Music and Migration. *Journal of Ethnic and Migration Studies*, 32(2):167–182, 2006.

J. G. A. Barbedo and A. Lopes. Automatic Musical Genre Classification Using a Flexible Approach. *Journal of Audio Engineering Society*, 56(7/8):560–568, 2008.

J. Barrett. World Music, nation and postcolonialism. *Cultural Studies*, 10(2): 237–247, 1996.

M. A. Bartsch and G. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, 2005.

G. F. Barz and T. J. Cooley, editors. *Shadows in the field: new perspectives for fieldwork in ethnomusicology*. Oxford University Press, 2008.

M. Bay, A. F. Ehmann, and J. S. Downie. Evaluation of multiple-f0 estimation and tracking sys- tems. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 315–320, 2009.

S. P. Bayard. Prolegomena to a Study of the Principal Melodic Families of British-American Folk Song. *The Journal of American Folklore*, 63(247):1–44, 1950.

B. Bel and B. Vecchione. Computational musicology. *Computers and the Humanities*, 27(1):1–5, 1993.

A. L. Berenzweig and D. P. Ellis. Locating singing voice segments within music signals. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 119–122, 2001.

J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

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

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

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

R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello. MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 155–160, 2014.

R. M. Bittner, J. Salamon, S. Essid, and J. P. Bello. Melody Extraction by Contour Classification. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 500–506, 2015.

R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello. Deep salience representations for F0 estimation in polyphonic music. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 63–70, 2017.

D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine learning (ICML)*, pages 113–120. ACM Press, 2006.

P. V. Bohlman. *World Music: A Very Short Introduction.* Oxford University Press, 2002.

B. Bozkurt. An Automatic Pitch Analysis Method for Turkish Maqam Music. *Journal of New Music Research*, 37(1):1–13, 2008. doi: 10.1080/09298210802259520.

B. H. Bronson. Mechanical Help in the Study of Folk Song. *Journal of American Folklore*, 62(244):81–86, 1949.

B. H. Bronson. Some Observations about Melodic Variation in British-American Folk Tunes. *Journal of the American Musicological Society*, 3:120–134, 1950.

B. H. Bronson. *The Traditional Tunes of the Child Ballads: With Their Texts, according to the Extant Records of Great Britain and America [4 Volumes].* Princeton University Press., Princeton, NJ, 1972.

S. Brown and J. Jordania. Universals in the world's musics. *Psychology of Music*, 41(2):229–248, 2011.

S. Brown, P. E. Savage, A. M.-S. Ko, M. Stoneking, Y.-C. Ko, J.-H. Loo, and J. A. Trejaut. Correlations in the population structure of music, genes and language. *Proceedings of the Royal Society B-Biological Sciences*, 281(1774), 2014. doi: 20132072.

J. A. Burgoyne, J. Wild, and I. Fujinaga. An Expert Ground-Truth Set for Audio Chord Recognition and Music Analysis. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 633–638, 2011.

J. A. Burgoyne, J. Wild, and I. Fujinaga. Compositional Data Analysis of Harmonic Structures in Popular Music. In J. Yust, J. Wild, and J. Burgoyne, editors, *Mathematics and Computation in Music. MCM 2013. Lecture Notes in Computer Science, vol 7937.* Springer, 2013.

M. J. Butler. *Unlocking the Groove.* Indiana University Press, Bloomington and Indianapolis, 2006.

J. J. Cabrera, J. M. Díaz-báñez, F. J. Escobar-Borrego, E. Gómez, F. Gómez, and J. Mora. Comparative Melodic Analysis of A Cappella Flamenco Cantes. In *Fourth Conference on Interdisciplinary Musicology (CIM08)*, pages 1–8, 2008.

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, 2008.

D. Chen, C. T. Lu, Y. Kou, and F. Chen. On detecting spatial outliers. *GeoInformatica*, 12(4):455–475, 2008.

J. Chen, S. Sathe, C. Aggarwal, and D. Turaga. Outlier Detection with Autoencoder Ensembles. In *Proceedings of the SIAM International Conference on Data Mining*, pages 90–98, 2017.

K. Choi, G. Fazekas, and M. Sandler. Automatic tagging using deep convolutional neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 805–811, 2016.

K. Choi, G. Fazekas, M. Sandler, and K. Cho. Transfer learning for music classification and retrieval tasks. In *International Society for Music Information Retrieval Conference*, pages 141–149, 2017a.

K. Choi, G. Fazekas, M. Sandler, and K. Cho. Convolutional recurrent neural networks for music classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2392–2396, 2017b.

P. Chordia and A. Rae. Raag recognition using pitch-class and pitch-class dyad distributions. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 431–436, 2007.

D. Clarke. On Not Losing Heart: A Response to Savage and Brown's "Toward a New Comparative Musicology". *Analytical Approaches to World Music*, 3 (2):1–14, 2014.

M. Clayton, T. Herbert, and R. Middleton, editors. *The cultural study of music: A critical introduction*. Routledge, New York, 2003.

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

D. Conklin and C. Anagnostopoulou. Comparative Pattern Analysis of Cretan Folk Songs. *Journal of New Music Research*, 40(2):119–125, 2011.

J. M. Cosell and L. LPCW: An LPC vocoder with linear predictive spectral warping. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 466–469, 1976.

E. Dahlig-Turek, S. Klotz, R. Parncutt, and F. Wiering. *Musicology (Re-) Mapped: Discussion Paper*. European Science Foundation, 2012. ISBN 9782918428848. URL `http://archives.esf.org/fileadmin/Public_documents/Publications/musicology.pdf`.

A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111 (4):1917–1930, 2002.

P. Desain, H. Honing, H. Vanthienen, and L. Windsor. Computational modeling of music cognition: problem or solution? *Music Perception: An Interdisciplinary Journal*, 16(1):151–166, 1998.

S. Dieleman and B. Schrauwen. Multiscale Approaches To Music Audio Feature Learning. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 116–121, 2013.

S. Dieleman and B. Schrauwen. End-to-end learning for music audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6964–6968, 2014.

S. Dixon, F. Gouyon, and G. Widmer. Towards Characterisation of Music via Rhythmic Patterns. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 509–516, 2004.

G. Dong and J. Li. Efficient Mining of Emerging Patterns: Discovering Trends and Differencesï£ij. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 43–52, 1999.

J. S. Downie. The Music Information Retrieval Evaluation Exchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

K. Dressler. An Auditory Streaming Approach for Melody Extraction from Polyphonic Music. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 19–24, 2011.

F. A. Dubinskas. A musical Joseph's coat: Patchwork patterns and social significance in world musics. *Reviews in Anthropology*, 10(3):27–42, 1983.

J. Durrieu, G. Richard, B. David, and C. FeÌĄvotte. Source/filter model for unsupervised main melody ex- traction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):564–575, 2010.

K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal. Speech/music discrimination for multimedia applications. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2445–2448, 2000.

T. M. Esparza, J. P. Bello, and E. J. Humphrey. From Genre Classification to Rhythm Similarity: Computational and Musicological Insights. *Journal of New Music Research*, 44(1):39–57, 2014.

S. Feld. Sound structure as social structure. *Ethnomusicology*, 28(3):383–409, 1984.

T. Fillon, J. Simonnot, M.-F. Mifune, S. Khoury, G. Pellerin, M. Le Coz, E. A. de la Bretèque, D. Doukhan, and D. Fourer. Telemeta: An open-source web framework for ethnomusicological audio archives management and automatic analysis. In *1st International Digital Libraries for Musicology workshop (DLfM 2014)*, pages 1–8, 2014. doi: 10.1145/2660168.2660169.

P. Filzmoser. A Multivariate Outlier Detection Method. In *International Conference on Computer Data Analysis and Modeling*, pages 18–22, 2004.

P. Filzmoser, R. Maronna, and M. Werner. Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, 52(3):1694–1711, 2008.

R. Fink. Big (Bad) Data, 2013. URL `http://musicologynow.ams-net.org/2013/08/big-bad-data.html`.

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.

D. Fourer, J.-l. Rouas, P. Hanna, and M. Robine. Automatic Timbre Classification of Ethnomusicological Audio Recordings. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 295–300, 2014.

R. Franzen. Europeana Sounds: an interface into European sound archives. *Sound Studies*, 2(1):103–106, 2016.

T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 464–467, 1999.

J. Futrelle and J. S. Downie. Interdisciplinary communities and research issues in Music Information Retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 215–221, 2002.

K. K. Ganguli, S. Gulati, X. Serra, and P. Rao. Data-Driven Exploration of Melodic Structures in Hindustani Music. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 605–611, 2016.

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

E. Gómez, M. Haro, and P. Herrera. Music and geography: Content description of musical audio from different parts of the world. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 753–758, 2009.

E. Gómez, P. Herrera, and F. Gómez-Martin. Computational Ethnomusicology: perspectives and challenges. *Journal of New Music Research*, 42(2):111–112, 2013.

A. J. Gustar. *Statistics in historical musicology*. PhD thesis, Open University, 2014.

P. Hamel and D. Eck. Learning Features from Music Audio with Deep Belief Networks. In *Proceedings of the International Society for Music Information Retrieval Conference*, number Ismir, pages 339–344, 2010.

H. Hammarström, R. Forkel, and M. Haspelmath. Glottolog 3.0, 2017. URL `http://glottolog.org/`.

M. Hammersley. Ethnography: problems and prospects. *Ethnography and Education*, 1(1):3–14, 2006.

P. Herrera and J. Bonada. Vibrato extraction and parameterization in the spectral modeling synthesis framework. In *Digital Audio Effects Conference*, pages 1–4, 1998.

W. Hess. *Pitch determination of speech signals: Algorithms and devices*. Springer-Verlag Berlin Heidelberg, 1983.

V. Hodge and J. Austin. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

A. Holzapfel. *Similarity methods for computational ethnomusicology*. PhD thesis, University of Crete, 2010.

A. Holzapfel and Y. Stylianou. Rhythmic Similarity in Traditional Turkish Music. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 99–104, 2009.

A. Holzapfel and Y. Stylianou. Scale Transform in Rhythmic Similarity of Music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):176–185, 2011.

A. Holzapfel, A. Flexer, and G. Widmer. Improving tempo-sensitive and tempo-robust descriptors for rhythmic similarity. In *Proceedings of the Sound and Music Computing Conference*, pages 247–252, 2011.

E. J. Humphrey, J. P. Bello, and Y. LeCun. Feature learning and deep architectures: New directions for music informatics. *Journal of Intelligent Information Systems*, 41(3):461–481, 2013.

D. Huron. The melodic arch in Western folksongs. *Computing in Musicology*, 10:3–23, 1996.

D. Huron. On the virtuous and the vexatious in an age of big data. *Music Perception*, 31(1):4–9, 2013.

C. Inskip and F. Wiering. In their own words: Using text analysis to identify musicologists' attitudes towards technology. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 455–461, 2015.

V. Ishwar, S. Dutta, A. Bellur, and H. A. Murthy. Motif Spotting in an Alapana in Carnatic Music. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 499–504, 2013.

S. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

Z. Juhász. A systematic comparison of different European folk music traditions using self-organizing maps. *Journal of New Music Research*, 35(2):95–112, 2006.

Z. Juhász. Automatic Segmentation and Comparative Study of Motives in Eleven Folk Song Collections using Self-Organizing Maps and Multidimensional Mapping. *Journal of New Music Research*, 38(1):71–85, 2009.

N. V. Kelso and T. Patterson. Natural Earth. URL `http://www.naturalearthdata.com`.

A. Klapuri, A. J. Eronen, and J. T. Astola. Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):342–355, 2006.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.

N. Kroher, E. Gómez, C. Guastavino, F. Gómez, and J. Bonada. Computational Models for Perceived Melodic SImilarity in A Capella Flamenco Singing. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 65–70, 2014.

N. Kroher, J.-M. Díaz-Báñez, J. Mora, and E. Gómez. Corpus COFLA: A Research Corpus for the Computational Study of Flamenco Music. *Journal on Computing and Cultural Heritage*, 9(2):10:1–10:21, 2016.

A. Kruspe, H. Lukashevich, J. Abeßer, H. Großmann, and C. Dittmar. Automatic Classification of Musical Pieces Into Global Cultural Areas. In *AES 42nd Conference: Semantic Audio*, 2011. URL `http://www.aes.org/e-lib/browse.cfm?elib=15958`.

O. Lartillot and P. Toiviainen. A Matlab Toolbox for Musical Feature Extraction From Audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

S. Le Bomin, G. Lecointre, and E. Heyer. The evolution of musical diversity: The key role of vertical transmission. *PLoS ONE*, 11(3), 2016.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

H. Lee, J. Yoo, and S. Choi. Semi-Supervised Nonnegative Matrix Factorization. *IEEE Signal Processing Letters*, 17(1):4–7, 2010.

M. Leman. Systematic musicology at the crossroads of modern music research. In A. Schneider, editor, *Systematic and Comparative Musicology: Concepts, Methods, Findings*, pages 89–115. Peter Lang, Frankfurt am Main, 2008.

P. H. Lindsay and D. A. Norman. *Human information processing: An introduction to psychology.* Academic Press, 1977.

B. Logan. Mel-Frequency Cepstral Coefficients for Music Modeling. In *Proceedings of the International Symposium on Music Information Retrieval*, 2000. doi: 10.1.1.11.9216.

A. Lomax. *Folk song style and culture*. American Association for the Advancement of Science, 1968.

A. Lomax. *Cantometrics: An Approach to the Anthropology of Music*. University of California Extension Media Center, Berkeley, 1976.

A. Lomax. Factors of musical style. In S. Diamond, editor, *Theory & practice: Essays presented to Gene Weltfish*, pages 29–58. Mouton, The Hague, 1980.

A. Lomax and N. Berkowitz. The Evolutionary Taxonomy of Culture. *Science*, 177(4045):228–239, 1972.

M. Marolt. A mid-level representation for melody-based retrieval in audio collections. *IEEE Transactions on Multimedia*, 10(8):1617–1625, 2008.

M. Marolt. Probabilistic Segmentation and Labeling of Ethnomusicological Field Recordings. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 75–80, 2009.

M. Marolt. Music/speech classification and detection submission for MIREX 2015. Technical report, 2015. URL `http://www.music-ir.org/mirex/abstracts/2015/MM3.pdf`.

S. Maruo, K. Yoshii, K. Itoyama, M. Mauch, and M. Goto. A feedback framework for improved chord recognition based on NMF-based approximate note transcription. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 196–200, 2015.

M. Mauch and S. Ewert. The Audio Degradation Toolbox and Its Application to Robustness Evaluation. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 83–88, 2013.

M. Mauch, K. Frieler, and S. Dixon. Intonation in Unaccompanied Singing : Accuracy , Drift and a Model of Reference Pitch Memory. *The Journal of the Acoustical Society of America*, 136(1):401–411, 2014.

M. Mauch, C. Cannam, R. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon. Computer-aided melody note transcription using the tony software: Accuracy and efficiency. In *Proceedings of the First International Conference on Technologies for Music Notation and Representation*, 2015a. doi: 10.1121/1.4881915.

M. Mauch, R. M. MacCallum, M. Levy, and A. M. Leroi. The evolution of popular music: USA 1960-2010. *Royal Society Open Science*, 2(5), 2015b. doi: 10.1098/rsos.150081.

B. McFee and J. P. Bello. Structured Training for Large-Vocabulary Chord Recognition. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 188–194, 2017.

B. McFee, E. Humphrey, and J. Bello. A software framework for Musical Data Augmentation. In *International Society for Music Information Retrival conference*, pages 248–254, 2015a.

B. McFee, M. McVicar, C. Raffel, D. Liang, O. Nieto, E. Battenberg, J. Moore, D. Ellis, R. Yamamoto, R. Bittner, D. Repetto, P. Viktorin, J. F. Santos, and A. Holovaty. librosa: 0.4.1, 2015b.

C. McKay. *Automatic music classification with jMIR*. Phd thesis, McGill University, Canada, 2010.

S. Mishra, B. Sturm, and S. Dixon. Local Interpretable Model-Agnostic Explanations for Music Content Analysis. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 537–543, 2017.

D. Moelants, O. Cornelis, and M. Leman. Exploring African Tone Scales. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 489–494, 2009.

J. Mora, F. Gómez, E. Gómez, F. Escobar-Borrego, and J. M. Díaz-Báñez. Characterization and Melodic Similarity of A Cappella Flamenco Cantes. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 351–356, 2010.

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.

Y. V. S. Murthy and S. G. Koolagudi. Classification of vocal and non-vocal regions from audio songs using spectral features and pitch variations. In *IEEE Canadian Conference on Electrical and Computer Engineering*, 2015. doi: 10.1109/CCECE.2015.7129461.

J. Nam, J. Herrera, M. Slaney, and J. Smith. Learning Sparse Feature Representations for Music Annotation and Retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 565–570, 2012.

J. Nam, J. Herrera, and K. Lee. A Deep Bag-of-Features Model for Music Auto-Tagging. Technical report, 2015. URL `http://arxiv.org/abs/1508.04999`.

N. Nettheim. A bibliography of statistical applications in musicology. *Musicology Australia*, 20(1):94–106, 1997.

B. Nettl. Review of Folk Song Style and Culture by Alan Lomax Source. *American Anthropologist, New Series*, 72(2):438–441, 1970.

B. Nettl. The Harmless Drudge: Defining Ethnomusicology. In *The Study of Ethnomusicology Thirty-one Issues and Concepts*, pages 3–15. University of Illinois Press, Urbana and Chicago, 2nd edition, 2005.

B. Nettl. Folk Music, 2014. URL `https://www.britannica.com/art/folk-music`.

B. Nettl. *The study of ethnomusicology: Thirty-three discussions*. University of Illinois Press, Champaign, 3rd ed. edition, 2015.

B. Nettl and P. V. Bohlman, editors. *Comparative musicology and anthropology of music: Essays on the history of ethnomusicology*. University of Chicago Press, Chicago, 1991.

B. Nettl, R. M. Stone, J. Porter, and T. Rice, editors. *The Garland Encyclopedia of World Music*. Garland Pub, New York, 1998-2002 edition, 1998.

K. Neubarth, M. Bergeron, and D. Conklin. Associations between musicology and music information retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 429–434, 2011.

Y. Ni, M. McVicar, R. Santos-Rodriguez, and T. D. Bie. An end-to-end machine learning system for harmonic analysis of music. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1771–1783, 2012.

S. Oramas, O. Nieto, F. Barbieri, and X. Serra. Multi-label music genre classification from audio, text, and images using deep features. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 23–27, 2017.

A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval. Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1564–1578, 2007.

F. Pachet and J.-J. Aucouturier. Improving timbre similarity: How high is the sky? *Journal of negative results in speech and audio sciences*, 1(1):1–13, 2004.

H. Pamjav, Z. Juhász, A. Zalán, E. Németh, and B. Damdin. A comparative phylogenetic study of genetics and folk music. *Molecular Genetics and Genomics*, 287(4):337–349, 2012.

E. Pampalk, A. Flexer, and G. Widmer. Improvements of Audio-Based Music Similarity and Genre Classification. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 634–637, 2005.

C. Panagiotakis and G. Tziritas. A speech/music discriminator based on RMS and zero-crossings. *IEEE Transactions on Multimedia*, 7(1):155–166, 2005.

M. Panteli and S. Dixon. On the evaluation of rhythmic and melodic descriptors for music similarity. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 468–474, 2016.

M. Panteli and H. Purwins. A Quantitative Comparison of Chrysanthine Theory and Performance Practice of Scale Tuning, Steps, and Prominence of the Octoechos in Byzantine Chant. *Journal of New Music Research*, 42(3):205–221, 2013.

M. Panteli, E. Benetos, and S. Dixon. Learning a feature space for similarity in world music. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 538–544, 2016a.

M. Panteli, E. Benetos, and S. Dixon. Automatic detection of outliers in world music collections. In *Analytical Approaches to World Music*, pages 1–4, 2016b.

M. Panteli, B. Rocha, N. Bogaards, and A. Honingh. A model for rhythm and timbre similarity in electronic dance music. *Musicae Scientiae*, 21(3):338–361, 2016c.

M. Panteli, R. Bittner, J. P. Bello, and S. Dixon. Towards the characterization of singing styles in world music. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 636–640, 2017.

G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *Technical Report. IRCAM*, 2004.

C. Pegg, H. Myers, P. V. Bohlman, and M. Stokes. Ethnomusicology. In S. Sadie, editor, *The New Grove Dictionary of Music and Musicians*. Macmillan, London, 2001.

T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer. On rhythm and general music similarity. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 525–530, 2009.

J. Pons and X. Serra. Designing efficient architectures for modeling temporal features with convolutional neural networks. In *International Conference on Acoustics, Speech and Signal Processing*, 2017.

A. Porter, M. Sordo, and X. Serra. Dunya: A System for Browsing Audio Music Collections Exploiting Cultural Context. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 101–106, 2013.

D. M. W. Powers. Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

E. Prame. Measurements of the vibrato rate of ten singers. *The journal of the Acoustical Society of America*, 96(4):1979–1984, 1994.

M. Prockup, A. F. Ehmann, F. Gouyon, E. M. Schmidt, O. Celma, and Y. E. Kim. Modeling genre with the Music Genome project: Comparing human-labeled attributes and audio features. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 31–37, 2015.

C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis. mir eval: A transparent imple- mentation of common mir metrics. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 367–372, 2014.

Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson. The Music Ontology. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 417–422, 2007.

W. Rhodes. The Use of Computer in the Classification of Folk Tunes. *Studia Musicologica*, VII:339–343, 1965.

F. Rodriguez Algarra, B. L. Sturm, and H. Maruri-Aguilar. Analysing Scattering-Based Music Content Analysis Systems: Where's the Music? In *International Society for Music Information Retrieval Conference*, pages 344–350, 2016.

S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.-L. Collette. Vibrato: detection, estimation, extraction, modification. In *Digital Audio Effects Workshop (DAFx'99)*, pages 1–4, 1999.

P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(C):53–65, 1987.

J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172, 2003.

T. Rzeszutek, P. E. Savage, and S. Brown. The structure of cross-cultural musical diversity. *Proceedings of the Royal Society B-Biological Sciences*, 279 (1733):1606–1612, 2012.

S. Sadie, J. Tyrrell, and M. Levy. *The New Grove Dictionary of Music and Musicians*. Oxford University Press, 2001.

J. Salamon and E. Gómez. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

J. Salamon, E. Gómez, and J. Bonada. Sinusoid extraction and salience function design for predominant melody estimation. In *International Conference on Digital Audio Effects*, pages 73–80, 2011.

J. Salamon, B. Rocha, and E. Gomez. Musical genre classification using melody features extracted from polyphonic music signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 81–84, 2012.

P. E. Savage. *Measuring the cultural evolution of music: With case studies of British-American and Japanese folk, art, and popular music*. Phd thesis, Tokyo University of the Arts, 2017.

P. E. Savage and S. Brown. Toward a new comparative musicology. *Analytical Approaches to World Music*, 2(2):148–197, 2013.

P. E. Savage and S. Brown. Mapping Music: Cluster Analysis Of Song-Type Frequencies Within and Between Cultures. *Ethnomusicology*, 58(1):133–155, 2014.

P. E. Savage, E. Merritt, T. Rzeszutek, and S. Brown. CantoCore: A new cross-cultural song classification scheme. *Analytical Approaches to World Music*, 2 (1):87–137, 2012.

P. E. Savage, S. Brown, E. Sakai, and T. E. Currie. Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences of the United States of America*, 112(29):8987–8992, 2015a.

P. E. Savage, H. Matsumae, H. Oota, M. Stoneking, T. E. Currie, A. Tajima, M. Gillan, and S. Brown. How "circumpolar" is Ainu music? Musical and genetic perspectives on the history of the Japanese archipelago. *Ethnomusicology Forum*, 24(3):443–467, 2015b.

H. Schaffrath. *The Essen Folksong Collection in the Humdrum Kern Format*. Center for Computer Assisted Research in the Humanities, Menlo Park, CA, 1995.

M. Schedl, E. Gomez, and J. Urbano. Music Information Retrieval: Recent Developments and Applications. *Foundations and Trends® in Information Retrieval*, 8(2-3):127–261, 2014.

E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1331–1334, 1997.

E. D. Scheirer. Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, 103(1):588–601, 1998.

E. G. Schellenberg and C. von Scheve. Emotional cues in American popular music: Five decades of the Top 40. *Psychology of Aesthetics, Creativity, and the Arts*, 6(3):196–203, 2012.

D. K. Scherrer and P. H. Scherrer. An Experiment in the Computer Measurement of Melodic Variation in Folksong. *The Journal of American Folklore*, 84 (332):230–241, 1971.

J. Schlüter. Learning to pinpoint singing voice from weakly labeled examples. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 44–50, 2016.

J. Schlüter and T. Grill. Exploring data augmentation for improved singing voice detection with neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 121–126, 2015.

D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer. Local and Global Scaling Reduce Hubs in Space. *Journal of Machine Learning Research*, 13:2871–2902, 2012.

J. Serrà, G. K. Koduri, M. Miron, and X. Serra. Assessing the tuning of sung indian classical music. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 157–162, 2011.

J. Serrà, Á. Corral, M. Boguñá, M. Haro, and J. L. Arcos. Measuring the Evolution of Contemporary Western Popular Music. *Scientific Reports*, 2 (521), 2012. doi: 10.1038/srep00521.

X. Serra. A Multicultural Approach in Music Information Research. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 151–156, 2011.

X. Serra. Creating research corpora for the computational study of music: the case of the CompMusic project. In *AES 53rd International Conference: Semantic Audio*, 2014. URL `http://www.aes.org/e-lib/browse.cfm?elib=17124`.

X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gómez, F. Gouyon, P. Herrera, S. Jorda, O. Paytuvi, G. Peeters, H. Schlüter, J., Vinet, and G. Widmer. *Roadmap for Music Information ReSearch*. Creative Commons BY-NC-ND 3.0 license., 2013.

K. Seyerlehner, M. Schedl, R. Sonnleitner, D. Hauger, and B. Ionescu. From Improved Auto-Taggers to Improved Music Similarity Measures. In A. Nürnberger, S. Stober, B. Larsen, and M. Detyniecki, editors, *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation. AMR 2012. Lecture Notes in Computer Science, vol 8382.*, pages 193–202. Springer, Cham, 2014.

U. Shalit, D. Weinshall, and G. Chechik. Modeling Musical Influence with Topic Models. In *Proceedings of the International Conference on Machine Learning*, pages 244–252, 2013.

D. Shanahan, K. Neubarth, and D. Conklin. Mining Musical Traits of Social Functions in Native American Music. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 681–687, 2016.

M. Slaney. Auditory Toolbox. Technical report, Interval Research Corporation, 1998.

A. South. rworldmap: A New R package for Mapping Global Data. *The R Journal*, 3(1):35–43, 2011.

A. Srinivasamurthy, A. Holzapfel, and X. Serra. In Search of Automatic Rhythm Analysis Methods for Turkish and Indian Art Music. *Journal of New Music Research*, 43(1):94–114, 2014.

C. J. Stevens. Music Perception and Cognition: A Review of Recent Cross-Cultural Research. *Topics in Cognitive Science*, 4:653–667, 2012.

S. Stevens and J. Volkmann. The Relation of Pitch to Frequency: A Revised Scale. *The American Journal of Psychology*, 53(3):329–353, 1940.

S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

S. Stober, D. J. Cameron, and J. A. Grahn. Classifying EEG recordings of rhythm perception. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 649–654, 2014.

B. L. Sturm. Classification accuracy is not enough. *Journal of Intelligent Information Systems*, 41(3):371–406, 2013.

B. L. Sturm. A Simple Method to Determine if a Music Information Retrieval System is a "Horse". *IEEE Transactions on Multimedia*, 16(6):1636–1644, 2014.

B. L. Sturm. Revisiting priorities: improving MIR evaluation practices. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 488–494, 2016.

L. Sun, S. Ji, and J. Ye. *Multi-Label Dimensionality Reduction*. CRC Press, Taylor & Francis Group, 2013.

D. Temperley and L. Van Handel. Introduction to the special issues on corpus methods. *Music Perception*, 31(1):1–3, 2013.

M. Tenzer, editor. *Analytical studies in world music*. Oxford University Press, New York, 2006.

D. Thompson. 1991: The Most Important Year in Pop-Music History. *The Atlantic*, 2015. URL `http://www.theatlantic.com/entertainment/archive/2015/05/1991-the-most-important-year-in-music/392642/`.

E. Thul and G. T. Toussaint. A Comparative Phylogenetic-Tree Analysis of African Timelines and North Indian Talas. In *Bridges Leeuwarden: Mathematics, Music, Art, Architecture, Culture*, pages 187–194, 2008.

J. T. Titon, T. J. Cooley, D. Locke, D. P. McAllester, and A. K. Rasmussen. *Worlds of Music: An Introduction to the Music of the World's Peoples*. Schirmer Cengage Learning, Belmont, 2009.

G. Toussaint. Classification and phylogenetic analysis of African ternary rhythm timelines. In *Meeting Alhambra, ISAMA-BRIDGES Conference*, pages 25–36, 2003.

S. E. Trehub. Cross-cultural convergence of musical features. *Proceedings of the National Academy of Sciences of the United States of America*, 112(29): 8809–8810, 2015.

G. Tzanetakis and P. Cook. MARSYAS: a framework for audio analysis. *Organised Sound*, 4(3):169–175, 2000.

G. Tzanetakis and P. Cook. Musical Genre Classification of Audio Signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

G. Tzanetakis, A. Kapur, A. W. Schloss, and M. Wright. Computational Ethnomusicology. *Journal of Interdisciplinary Music Studies*, 1(2):1–24, 2007.

T. Underwood. Can we date revolutions in the history of literature and music?, 2015. URL `http://tedunderwood.com/2015/10/03/can-we-date-revolutions-in-the-history-of-literature-and-music/`.

T. Underwood, H. Long, R. J. So, and Y. Zhu. You say you found a revolution, 2016. URL `http://tedunderwood.com/2016/02/07/you-say-you-found-a-revolution/`.

J. Urbano, D. Bogdanov, P. Herrera, E. Gómez, and X. Serra. What is the effect of audio quality on the robustness of MFCCs and chroma features. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 573–578, 2014.

J. Van Balen, D. Bountouridis, F. Wiering, and R. Veltkamp. Cognition-inspired Descriptors for Scalable Cover Song Retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 379–384, 2014.

A. Van Den Oord, S. Dieleman, and B. Schrauwen. Transfer learning by supervised pre-training for audio-based music classification. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 29–34, 2014.

L. van der Maaten and G. Hinton. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

P. van Kranenburg, J. Garbers, A. Volk, F. Wiering, L. Grijp, and R. C. Veltkamp. Collaborative perspectives for folk song research and music information retrieval: The indispensable role of computational musicology. *Journal of Interdisciplinary Music Studies*, 4(1):17–43, 2010.

P. van Kranenburg, A. Volk, and F. Wiering. A Comparison between Global and Local Features for Computational Classification of Folk Song Melodies. *Journal of New Music Research*, 42(1):1–18, 2013.

P. van Kranenburg, M. de Bruin, L. P. Grijp, and F. Wiering. *The Meertens Tune Collections*. Meertens Institute, Amsterdam, 1st edition, 2014.

V. Viro. Peachnote: Music score search and analysis platform. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 359– 362, 2011.

A. Volk and W. B. de Haas. A Corpus-Based Study on Ragtime Syncopation. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 163–168, 2013.

A. Volk and P. van Kranenburg. Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae*, 16(3):317–339, 2012.

A. Volk, F. Wiering, and P. van Kranenburg. Unfolding the Potential of Computational Musicology. In *Proceedings of the13th International Conference on Informatics and Semiotics in Organisations (ICISO)*, pages 137–144, 2011.

E. M. von Hornbostel and C. Sachs. Classification of musical instruments. *Galpin Society Journal*, 14:3–29, 1961.

Z. Wallmark. Big Data and Musicology: New Methods, New Questions, 2013. URL `http://www.academia.edu/6442281/Big_Data_and_Musicology_New_Methods_New_Questions`.

C. Walshaw and C. Walshaw. A Statistical Analysis of the ABC Music Notation Corpus : Exploring Duplication. In *Proceedings of the Fourth International Workshop on Folk Music Analysis*, 2014. doi: 10.13140/2.1.4340.0961.

T. C. Walters, D. A. Ross, and R. F. Lyon. The Intervalgram: An Audio Feature for Large-scale Melody Recognition. In *9th International Symposium on Computer Music Modeling and Retrieval*, pages 19–22, 2012.

F. Wiering and E. Benetos. Digital Musicology and MIR : Papers , Projects and Challenges. In *ISMIR 2013 Late-breaking session*, 2013. URL `http://ismir2013.ismir.net/wp-content/uploads/2014/02/lbd4.pdf`.

G. A. Wiggins. Semantic Gap?? Schemantic Schmap!! Methodological Considerations in the Scientific Study of Music. In *IEEE International Symposium on Multimedia*, 2009. doi: 10.1109/ISM.2009.36.

D. Wolff and T. Weyde. Adapting Metrics for Music Similarity Using Comparative Ratings. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 73–78, 2011.

F. Zhou, Q. Claire, and R. D. King. Predicting the Geographical Origin of Music. In *IEEE International Conference on Data Mining*, pages 1115–1120, 2014.

P. H. R. Zivic, F. Shifres, and G. A. Cecchi. Perceptual basis of evolving Western musical styles. *Proceedings of the National Academy of Sciences of the United States of America*, 110(24):10034–10038, 2013.

S. Åđentürk, S. Gulati, and X. Serra. Score informed tonic identification for makam music of Turkey. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 175–180, 2013.