AUDIO DESCRIPTION AND CORPUS ANALYSIS OF POPULAR MUSIC

Beschrijving en corpusanalyse van populaire muziek
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op woensdag 15 juni 2016 des middags te 2.30 uur

door

JAN MARKUS HARRIE VAN BALEN

geboren op 26 oktober 1988
te Leuven, België

Promotor: Prof.dr. R.C. Veltkamp
Copromotor: Dr. F. Wiering

The research in this thesis was supported by the NWO CATCH project COGITCH (640.005.004).

ABSTRACT

In the field of sound and music computing, only a handful of studies are concerned with the pursuit of new musical knowledge. There is a substantial body of *corpus analysis* research focused on new musical insight, but almost all of it deals with symbolic data: scores, chords or manual annotations. In contrast, and despite the wide availability of audio data and tools for audio content analysis, very little work has been done on the corpus analysis of audio data.

This thesis presents a number of contributions to the scientific study of music, based on *audio corpus analysis*. We focus on three themes: audio description, corpus analysis methodology, and the application of these description and analysis techniques to the study of music similarity and 'hooks'.

On the theme of audio description, we first present, in part i, an overview of the audio description methods that have been proposed in the music information retrieval literature, focusing on timbre, harmony and melody. We critically review current practices in terms of their relevancy to audio corpus analysis. Throughout part ii and iii, we then propose new feature sets and audio description strategies. Contributions include the introduction of *audio bigram* features, pitch descriptors that can be used for retrieval as well as corpus analysis, and *second-order audio features*, which quantify distinctiveness and recurrence of feature values given a reference corpus.

On the theme of audio corpus analysis methodology, we first situate corpus analysis in the disciplinary context of music information retrieval, empirical musicology and music cognition. In part i, we then present a review of audio corpus analysis, and a case study comparing two influential corpus-based investigations into the evolution of popular music [122, 175]. Based on this analysis, we formulate a set of nine recommendations for audio corpus analysis research. In part ii and iii, we present, alongside the new audio description techniques,

new analysis methods for the study of song sections and within-song variation in a large corpus. Contributions on this theme include the first use of a probabilistic graphical model for the analysis of audio features.

Finally, we apply new audio description and corpus analysis techniques to address two research problems of the COGITCH project of which our research was a part: improving audio-based models of music similarity, and the analysis of hooks in popular music. In parts i and ii, we introduce *soft audio fingerprinting*, an umbrella MIR task that includes any efficient audio-based content identification. We then focus on the problem of scalable cover song detection, and evaluate several solutions based on audio bigram features. In part iii, we review the prevailing perspectives on musical catchiness, recognisability and hooks. We describe Hooked, a game we designed to collect data on the recognisability of a set of song fragments. We then present a corpus analysis of hooks, and new findings on what makes music catchy.

Across the three themes above, we present several contributions to the available methods and technologies for audio description and audio corpus analysis. Along the way, we present new insights into choruses, catchiness, recognisability and hooks. By applying the proposed technologies, following the proposed methods, we show that rigorous audio corpus analysis is possible and that the technologies to engage in it are available.

ACKNOWLEDGEMENTS

For the last four years, I have been able to do work I thoroughly enjoy on a topic I truly care about, and I have been able to do this because of the vision and work of everyone involved in making the COGITCH project a reality. I am very grateful to my supervisors Frans and Remco for their guidance, feedback and ideas. Your help has been essential to this thesis, and the freedom you have given me to explore many different topics from many different angles has helped me find my passion as a researcher.

I also want to thank my closest collaborators on this project, Dimitrios and Ashley, for all the amazing work they did, and for the many discussions we had at the office. This research would have been far less exciting without the amazing dataset that is the *Hooked!* and *Hooked on Music* data. It wouldn't have existed without their tireless work on both of the games. It also wouldn't have existed without the work of Henkjan Honing. I am grateful for his initiative and guidance, and for making me feel welcome in the Music Cognition Group and music reading group at the University of Amsterdam.

I'm also grateful for many insightful discussions with colleages at the Meertens Institute and Beeld en Geluid. I especially thank everyone at the Meertens Institute for giving me a place to work in Amsterdam, and to the members of Tunes and Tales and FACT for their company. A special word of gratitude goes out to Louis Grijp. Louis' extraordinary vision laid the foundations for much of the digital music research in The Netherlands, and the COGITCH project would not have existed without his decades of work on the Dutch Song Database.

For their support inside and outside the office, I also thank my fellow PhD students in the department—Marcelo, Vincent, Anna—as well as in Amsterdam, London and Barcelona, and at the different ISMIR conferences I was lucky to attend; I learned a lot from the interactions with students of my generation.

Finally, I would like to thank my brothers and my parents for their patience, hospitality and help, you gave me a welcome home in Belgium, London and South Africa. A very big thank you also goes to Maria for her patience and support—and letting me turn our house into a coffee lab during the last months. And last but not least, I must thank my friends in Leuven and Amsterdam: thank you for the drinks and dinners that kept me sane.

CONTENTS

ACRONYMS

ACF	Autocorrelation function
DTW	Dynamic time warping
DFT	Discrete Fourier transform
FFT	Fast Fourier transform
GMM	Gaussian mixture mode
HMM	Hidden Markov model
HPCP	Harmonic pitch class profile
KNN	k nearest neighbours
LDA	Latent Dirichlet allocation
LMER	Linear mixed effects regression
LSA	Latent semantic analysis
MAP	Mean average precision
MER	Music emotion recognition
MFCC	Mel frequency cepstral coefficients
MGR	Music genre recognition
MIREX	Music Information Retrieval Evaluation eXchange
MIR	Music information retrieval
MSD	Million Song Dataset
NMF	Non-negative matrix factorisation
OMR	Optical music recognition
PCA	Principal components analysis
PGM	Probabilistic graphical model
PLSR	partial least squares regression
SDM	Self-distance matrix
SM	Similarity matrix
SSM	Self-similarity matrix
STFT	Short-time Fourier transform

RELATED PUBLICATIONS

CHAPTER 4

Van Balen, J. (2013). A Computational Study of Choruses in Early Dutch Popular Music. In *Third International Workshop on Folk Music Analysis (FMA)*. Amsterdam, the Netherlands.

Van Balen, J., Burgoyne, J. A., Wiering, F., & Veltkamp, R. C. (2013). An Analysis of Chorus Features in Popular Song. In *14th International Society for Music Information Retrieval Conference (ISMIR)*. Curitiba, Brazil.

CHAPTER 5

Van Balen, J., Wiering, F., & Veltkamp, R. (2014). Cognitive Features for Cover Song Retrieval and Analysis. In *Fourth International Workshop on Folk Music Analysis (FMA)*. Istanbul, Turkey.

Van Balen, J., Bountouridis, D., Wiering, F., & Veltkamp, R. (2014). Cognition-inspired Descriptors for Scalable Cover Song Retrieval. In *15th International Society for Music Information Retrieval Conference (ISMIR)*. Taipe, Taiwan.

Bountouridis, D., & Van Balen, J. (2014). Towards Capturing Melodic Stability. In *Conference on Interdisciplinary Musicology*. Berlin, Germany.

CHAPTER 6

Van Balen, J., Wiering, F., & Veltkamp, R. (2015). Audio Bigrams as a Unifying Model of Pitch-based Song Description. In *11th International*

Symposium on Computer Music Multidisciplinary Research (CMMR). Plymouth, United Kingdom.

CHAPTER 7

Burgoyne, J. A., Bountouridis, D., Van Balen, J., & Honing, H. (2013). Hooked: a Game for Discovering What Makes Music Catchy. In *14th International Society for Music Information Retrieval Conference (ISMIR)*. Curitiba, Brazil.

Aljanaki, A., Bountouridis, D., Ashley, J., Wiering, F., Van Balen, J., & Honing, H. (2013). Designing Games with a Purpose for Music Research: Two Case Studies. In *The Games And Learning Alliance Conference (GALA)*. Paris, France.

Burgoyne, J. A., Van Balen, J., Bountouridis, D., Karavellas, T., Wiering, F., Veltkamp, R. C., & Honing, H. (2014). The Contours of Catchiness, or Where to Look for a Hook. In *International Conference on Music Perception and Cognition (ICMPC)*. Seoul, South Korea.

CHAPTER 8

Van Balen, J., Burgoyne, J. A., Bountouridis, D., MÏlensiefen, D., & Veltkamp, R. (2015). Corpus Analysis Tools for Computational Hook Discovery. In *16th International Society for Music Information Retrieval Conference (ISMIR)*. Malaga, Spain.

Part I

INTRODUCTION TO AUDIO FEATURES AND CORPUS ANALYSIS

1

INTRODUCTION

1.1 RESEARCH GOALS

1.1.1 *Audio Corpus Analysis*

The last decade saw the rapid growth of the digital humanities, an interdisciplinary area of research in which research topics and methods from the humanities and computing come together. The rise of digital humanities research can be explained by the unprecedented availability of tools and resources for *data-intensive* research. In linguistics, for example, it is now easier than ever to evaluate the evidence for a theory or hypothesis not just in a small selection of documents, but in a large corpus. Digital linguists, musicologists and other humanists owe this opportunity to pioneering efforts that go back to the beginning of the computing era, including digitization programs, the creation of various data formats and the developments of new infrastructures for off line and on line data.

Musicology has seen decades of digital and computational research, beginning as early as the 1960's and 1970's. In the late 1990's, building on developments in early computational musicology, digital signal processing and the web, developments in the field of music information retrieval (MIR) have given music research a digital boost from another, more consumer-oriented angle: that of music search and recommendation. MIR's continued pursuit of new data analysis methods has provided the music research community with a huge array of methods for the quantitative analysis of music, at unprecedented scale.

Over the last years, researchers have turned to these technologies to engage in ever more complex, large-scale and data-intensive music analysis—researchers with diverse backgrounds in both statistical or 'database-driven' musicology, and music information retrieval. *Corpus analysis* has been used in the search for musical 'universals' (universal properties of music) [171], to find and track trends in a sample of musical works, e.g., Western classical compositions [166] and Western popular music [42, 122, 175], or to model 'stability' of musical motives under oral transmission [197]. Corpus-level music analysis has also been used to test theories of expectation [80], or correlate features of the music with performance on various tasks, ranging from a memory test to walking [104].

A large majority of this research deals with *symbolic* data: scores, chords or manual annotations. This is not surprising given the origins of this research in computational musicology, but it contrasts sharply with the predominance, in music information retrieval, of research on *audio*, i.e., music recordings. Despite the wide availability of audio data and tools for audio content analysis, very little work has been done on the corpus analysis of audio data.

This thesis presents a number of contributions to the scientific study of music based on *audio corpus analysis*. We will begin this investigation with a closer look at what audio corpus analysis is, and how it fits into the larger context of music research, reviewing the fields of research of which it is a part, and laying out the argument for a corpus analysis based on audio, rather than symbolic data (section 1.2). But first, in section 1.1.2, we introduce the COGITCH project, the initiative behind this research, and discuss its objectives, to motivate the research goals of this thesis.

1.1.2 *The* COGITCH *project*

The COGITCH project was part of CATCH, a Netherlands-based science program for research on the intersection of cultural heritage and information technology, and financed by NWO, the Dutch organisation for scientific research. The COGITCH project was a collaboration between

two heritage institutes and two universities. On the side of the heritage institutes, the Meertens Institute (MI) is involved in the research and documentation of Dutch language and culture, and the Netherlands Institute for Sound and Vision (S&V) oversees, among other things, an archive of Dutch media heritage, including music. The research groups affiliated with the universities are the department of information and computing sciences of Utrecht University and the music cognition group at the University of Amsterdam.[1]

Both MI's and S&V's collections are very rich in data: the MI's *Dutch Song Database* contains metadata for over 140,000 songs, and audio recordings for a subset of them, including 7178 unique field recordings of Dutch folk songs [98].[2] The S&V collection contains metadata and audio for over 300,000 songs which it rents out to various media institutions, and an additional, physical collection of over 50,000 vinyl records (33, 45 and 78 RPM), part of which was digitized during the COGITCH project.

Access to Digital Music Heritage

The COGITCH project focused on two main objectives. The first goal of the COGITCH project was to facilitate access to both institutions' music collections, through an integrated search infrastructure, making the collections *interoperable*. Inspired by the technologies of an earlier CATCH project, WITCHCRAFT, this interoperability should extend into not just the metadata, but also the content of the collections: the music recordings.[3]

As will be explained in chapter 5, content-based retrieval within and between these collections requires a model of *similarity* between documents in the collection, which in turn requires powerful fragment-

1 http://www.uu.nl/organisatie/departement-informatica
http://mcg.uva.nl

2 http://www.liederenbank.nl/index.php?lan=en

3 The goal of the WITCHCRAFT project (2006-2010) was to create a 'functional content-based retrieval system for folksong melodies' using the Meertens Institute's collection of symbolic folk song transcriptions.
http://www.cs.uu.nl/research/projects/witchcraft/

level audio description methods and scalable methods for the comparison of these descriptions. Apart from being a useful advancement in itself, a good model of similarity for the documents in these collections can also benefit research into the evolution of folk song music in The Netherlands, the dynamics of stability and variation in oral traditions, and the emergence of popular music in the twentieth century [196].

Hooks and Memory in Popular Music

A second, equally important goal of the COGITCH project was to establish a scientific model of *hooks*. A hook, as will be explained in chapter 7, can be defined as the part of a song that is most recognizable. With a scientific investigation into hooks and the recognizability of real-world music, the COGITCH project seeks to improve our understanding of the role of memory in popular music. Eventually, we hope this will contribute to a better understanding of music memory in general.

A model of hooks can also help create better, more perceptually and cognitively informed similarity models and retrieval methods. Part of this second goal of the COGITCH project was to use any findings on hooks and memory to improve the similarity models discussed above and inform the aforementioned investigations into stability, variation and the evolution of folk and popular music. But it can be used to support other kinds of information retrieval, too: search, recommendation and browsing systems can be improved with better models of what users may or may not remember well [69].

1.1.3 *Research Goals*

The goals of this thesis are threefold. The first two relate to audio corpus analysis. First, we aim to review and advance the available audio description methods for corpus analysis research. Second, we aim to review the corpus analysis methodology itself, and explore new methods that address some of the open issues.

The third goal relates to the objectives of the COGITCH project. Not all of these will be addressed: some of the project goals that fall outside the scope of this thesis are the analysis of stability and variation, and the application of new findings on hooks and music memory to improve music retrieval technologies. The two objectives of the project that will be discussed in this thesis are the application of new audio description methods to develop new approaches to scalable song similarity, and a computational analysis of hooks.

These goals will not be addressed in the order they are now stated. Section 1.3, the last section of this chapter, will present an outline of the remainder of this thesis, with a focus on how the above objectives will be approached.

1.2 DISCIPLINARY CONTEXT

What is corpus analysis, and where does it fit in the larger context of music research? Corpus analysis was described above as 'data-intensive music research'. A more specific definition that we shall use in this thesis is: the analysis of a music collection with the aim of gaining insights into the music. Not any collection, of course: given one or more research questions, a dataset should be selected to represent the particular music to which the questions pertain.

The above definition seems to 'exclude' a sizable segment of computational and digital music research. Indeed, not all digital musicology involves the content-based analysis of a music collection. Conversely, most MIR research does not aim at new insights into music. To see this, we need to take a closer look at what musicology is, what MIR is, and how research in musicology and MIR is done. At the end of this discussion we also introduce *empirical musicology* and *cognitive musicology* and look at how they relate to musicology and MIR. Finally, we discuss how corpus analysis research fits in.

1.2.1 *Musicology in the Twentieth Century.*

Summarising a century of research in just a few paragraphs necessarily involves oversimplifications. However, it is fair to say that music scholarship, at the beginning of the twentieth century, had a strong focus on individual musical works as represented by the score. A lot of work was done recovering, editing and analyzing music sources, and much of it was concerned with the music that was at that time seen as Europe's great works of art, identifying the particular structural, musical aspects of these compositions that made them into the masterpieces for which they were seen. For many, the aim was to expose the nature of the true and the beautiful, so as to advance the art itself. Mastership, genius and beauty were seen as absolute, and in the perspective of the romantic ideals of the time, were to be found in instrumental music above all, and particularly, in its use of harmony [36].

This 'analysis'-centered musicology continued further into the twentieth century, with a positivist approach to music analysis which assumed that, like in the study of the physical world, objective laws could be found that underlie the way art works. Proponents of this approach, like Schoenberg and Schenker, pushed musicologists towards formalisation, establishing a theory of music that emphasized structure, abstraction and rules. Similar principles would inform the first endeavours into computational music analysis, which included attempts to implement these positivist approaches to music analysis, including 'Schenkerian analysis', as a computer program (e.g., [87]).

As musicology evolved, however, scholars started to see problems with the positivist project. One way this manifested, is through a shift away from the emphasis on music as autonomous 'works' to a kind of process, in which the performer and the listener play an important role, too. The shift was partially pioneered by scholars of non-Western music, who, much before music theorists and historical musicologists, broke with the positivist traditions from before the second World War. Particularly, 'comparative musicology', which aimed to understand the causes and mechanisms that shape the evolution of music across

cultures, was replaced with ethnomusicology, a new field based on the paradigms of cultural anthropology [170].

Nonetheless, a strong association persisted between the 'chronology' of the musical production chain, in which composition precedes performance and listening , and a hierarchy of prestige with composition ranking above performance and listening. A stronger paradigm shift came with the arrival of 'new musicology' around 1985 [88]. Looking back at the nineteenth century, and how little had changed up until the 1950's, musicologists became increasingly critical of the idea that there is such a thing as objective beauty or greatness. They also pointed to other perceived flaws in the positivist program, such as the assumption, in the search for a single authoritative reading for every musical work, that it was possible to figure out a composer's intentions. These intentions cannot be known, the new musicologists argued, and, more fundamentally, neither the intentions or the music can be understood onto itself. Music is a medium that influences and is influenced by feelings, desires and societal context such as power structures and taboo [36].

In this light, much of the historical writings on music were seen as justifications of the canon of a certain time, and the canon itself a product of the consensus of a cultural and political elite. The scholarship of the nineteenth century was therefore complicit in providing the justification for the cultural and political power structures of their time. Influenced by critical theory, feminism and gender studies, a new 'critical' musicology emerged, with the intention to understand music as it interacts with society and expose ideologies in music and music writing (e.g., [124]).

How did the proponents of computing respond to this? Judging by scientific output, computational musicology, as it is now referred to, all but disappeared in the 1980's [22, 201]. After the first initiatives mentioned at the very beginning of this chapter, a lot of effort was spent keeping up with ever-changing computing architectures (from mainframes to personal computers) and storage formats (from punch cards to floppies and hard drives). Meanwhile, the paradigm shift of new musicology tempered the heroic ambitions of these endeavours,

and music computation, as an agenda, became unfashionable. A new wave of digital music research, however, came with the emergence of music information retrieval in the late 1990's.

1.2.2 *Music Information Retrieval*

Music information retrieval (MIR) is typically described as an interdisciplinary field of scientific research, with origins in computing, library science and musicology [47]. The roots of MIR largely overlapped with the early work in computational musicology so, like computational musicology, it all but disappeared in the 1980's.

In the 1990's, however, digital audio became more widely available, and computing power surged. Music information retrieval revived as an area of *applied* research, with a partial re-orientation to audio-based research, but most of all, a strong focus on *tasks*, in which specific kinds of information are extracted from musical data [22]. In the task of optical music recognition (OMR), for example, a computer program is given, as input, an image of a score, and outputs a digital version of the score in some symbolic format.

The kind of tasks at the forefront of MIR evolved throughout the 1990's and 2000's, to include input data such as images, symbolic data, digital audio, metadata and crowd-sourced social tags [22]. Types of output that MIR systems are engineered to produce can roughly be divided into: metadata (as in recommendation systems), classification labels (as in genre classification or key finding), and symbolic sequences (as in chord labeling and other kinds of transcription) [22]. Some of these tasks will be presented in more detail in chapter 2.

The most important MIR development that came in the 2000's, was probably the introduction of MIREX, an organisational framework for the joint evaluation of new MIR algorithms under rigorous tests conditions, using common datasets and evaluation scripts [47]. Growing rapidly since its first edition in 2005, MIREX now annually evaluates the algorithms of over 100 researchers. This helped identify and advance the state of the art across 24 MIR tasks.

Music information retrieval technologies not only have become a part of everyday life (as they are incorporated in, e.g., music streaming services), they have the potential to become powerful tools in musicology. But did the progress made by MIR address the concerns of new musicology better than the first efforts in computational music research?

An optimist's evaluation would probably point to the progress made in scholars' access to digital resources. Optical music recognition tools are a relevant example, as well as on line music libraries and music typesetting tools such as Finale and Sibelius, which already form an essential part of many musicologists professional 'work flow' [83]. In other word, access to data has further improved since the day of first digital encoding projects. In itself however, this facilitation of access doesn't fundamentally address any of the new musicologists' concerns.

The pessimist might conclude that MIR has moved away from musicology entirely: the success of MIREX has made evaluation such a central part of MIR that it is now only focused on the kinds of analysis for which a *ground truth* exists. And whether such a ground truth really exist is a subject of continued debate. In other words, the 'computability paradigm' of early computational music research, which treats empirical data as a hermetic ground truth, and the 'critical' perspectives of new musicology, in which any ground truth is necessarily based on assumptions and hypotheses, have not been truly reconciled.

Furthermore, this focus on tasks with a 'ground truth' may be hampering the development of methods for analyzing music data to discover something 'new'. Burgoyne et al., in their review of MIR for a new digital humanities companion, conclude: "As soon as computers became a part of the academic infrastructure, researchers became interested in using them to study music. Over a period of some decades, the computers have gotten better at answering research questions" [22]. What we argue here is that perhaps computers have not, in fact, worked as much on answering research questions as on 'solving tasks'. Research questions have centered largely on issues like: 'can new method X be used to increase performance on task Y?', and as

a result, computers got better at Y. It may be time to put these technologies to work in answering actual research questions, not about method X or task Y, but about the music.

1.2.3 *Empirical Musicology and Music Cognition*

While music information retrieval branched off and diverged from musicology itself, empirical and quantitative methods in music research revived in another way. The revival coincided with a renewed support for empirical methods, a 'new empiricism', that was not just limited to music research. The new empiricism gave rise to a new *empirical musicology* (sometimes *systematic musicology*, an older and broader term mostly used in continental Europe), of which empiricism, formalisation and computation were an important part. And last but not least, both developments also propelled the new research area of *cognitive musicology* or *music cognition*. We now discuss this chain of developments, beginning with Huron's analysis of the new empiricism in [79].

Empirical Musicology

Huron first examines the new musicologists' resistance to empirical methods. New musicology, and postmodernism in general, tend to assume that there is no absolute truth to be known. Instead, truth is seen as a "social construction that relates to a local or partial perspective on the world". In other words, there is no privileged perspective or interpretation, and postmodernists are right to point this out.

However, empiricism and postmodernism are also very similar in a different way: both "cultivate institutionalized forms of skepticism". The kind of skepticism typically associated with the sciences involves holding scientific claims up to a standard of evidence that is focused on minimizing the number of 'false positive' claims, i.e., claims that are accepted as true even though they are not (strong emphasis on p-values, for example, reflects this focus). For many in arts and humanities research, however, a common fear is to make 'false negative' errors, to dismiss a claim that might in fact have merit—assigning

claims such certainty that they provide explanatory "closure", may be regarded as a provocation, "a political act intended to usurp all other views" [79]. The difference may be due to the level of risk associated with each kind of error (a 'false positive' claim in the humanities, for example, may be considered as relatively harmless compared to a false positive in the sciences) and with the amount of data typically available to test claims. In other words, the humanities and the sciences "might diverge in their philosophical conceptions about the nature of the world, they nevertheless share deep methodological commonalities" [79].

Furthermore, "even if we accept the proposition that there is no privileged interpretation, it does not necessarily follow that all interpretations are equally valid", or else all knowledge would be impossible. Looking at it from a cognitive angle, Huron notes: while "we should recognize that human beings are cultural entities, we must recognize that humans are also biological entities with a priori instinctive and dispositional knowledge about the world, that originates in an inductive process of evolutionary adaptation" [79].

Huron concludes that not all forms of rigor and empiricism should be abandoned, provided that we have data and a strategy to appropriately balance type-I and type-II errors, and provided that we navigate the 'known potholes' associated with the methodologies of choice, such as logical and rhetorical fallacies and statistical self-deception.

Honing, in 2004, also looked at the comeback of empiricism in musicology, and observed three trends. The first trend was the emergence of a revitalized systematic musicology, that "is based on empirical observation and rigorous method, but at the same time is also aware of, and accounts for, the social and cultural context in which music functions" [68]. A reconnection was found between the empiricists and the new musicologists, much in the way Huron described in [79].

The second trend Honing includes is the growing role of formalisation and computation in musicology, discussed earlier. This is a trend that precedes Huron's 'new empiricism'. Did these computational approaches see a similar adaptation to the concerns of postmodernism

(like the new empirical method), or not (like MIR)? Notably, Huron in [79] speaks of quantitative methods, but not of computation.

A classic example of formalisation in music theory is Lerdahl and Jackendoff's 'Generative Theory of Tonal Music' (GTTM), a highly formalized attempt to model the cognitive processing of Western tonal music, inspired in part by the proto-generative theory of Schenker and the generative grammars of Chomsky [105, 152]. The model was a landmark in the coming of age of music cognition and heavily influenced important interdisciplinary music research. But it has also been criticized, e.g., for not being a theory in the scientific sense of the word (i.e., "subject to testing and potentially falsification by hypothesis formation and experiment") and for treating music, like language in the work of Chomsky, as some kind of external absolute: "observing an idealised version of the phenomenon, and treating it as though it had its own existence" [37, 203]. To the new empiricist, in other words, it is neither empirical, nor is it much informed by the postmodernists' critique of music as an object. Another example given by Wiggins is Temperley's Grouper algorithm, a computational model for the segmentation of musical phrases [203]. In this case, the algorithm relies on the *ad-hoc* setting of a parameter to produce plausible segmentations for a particular style of music. To Wiggins, this makes it effectively a descriptive, rather than predictive theory. Nonetheless, descriptive models such as GTTM and Grouper can be a stepping stone to make progress towards a more explanatory, prescriptive model, and computational algorithms also get the merit of having given a greater visibility to of musicology outside the humanities [68, 203].

Music Cognition

Parallel to these new empirical and computational trends in musicology, was the 'cognitive revolution', Honing's third trend, marked by a similar interest in both rigorous empiricism, and computation [68].

The cognitive sciences are concerned with various aspects of the mind including perception, attention, memory, language, action, and emotion. Theories on how the mind works go back to Ancient Greece.

Only since the 1970's, however, 'cognitive science' was recognized as a research area of its own, with its own name and its own agenda, driven by twentieth century advances in the theory of computation (e.g., the work of Alan Turing), linguistics (e.g., Chomsky) artificial intelligence (Marvin Minsky and others) and other fields [152]. The emergence of cognitive science as a field of research created a new home for the study of a variety of aspects of music, too, including perception, expectation, music memory and emotion.

Music cognition is most distinct from twentieth century musicology, in that it regards music as fundamentally a psychological entity. As such, it enabled a use of empiricism and computing in music research that had not been seriously considered before: as scientific theories of music listening and performance. Wiggins et al., for example, argue for "a theory of music which starts from the position that music is primarily a construct of human minds" [203].

From early on, computational modeling became an important tool in the cognitive sciences. Honing recounts in 2011: "there is hardly a cognitive theory without a computational component" [70]. Computational modeling is distinct from other empirical methods in that it is not an instrument for quantitative observation. It is a methodological cycle that integrates theory and observation. True computational models are precisely formalized theories, therefore, like theories, they generate new hypotheses, which can in turn be tested empirically.

Summary

To summarize, some of the work done in musicology and computational musicology hasn't necessarily been striking the right balance between empiricism and the concerns of new musicology. However, an empirical approach to music research is possible that acknowledges these concerns, as laid out by Huron. A particularly promising place to look for a connection between an empirical methodology and a critical perspective of music is in cognitive musicology, which balances these concerns by emphasizing the need of predictive models and an approach to music from the perspective of the listener. By research-

ing music at the corpus level, we can aspire to a research method that accounts for the both the context and cognition of listening, to acknowledge that music has a cultural and cognitive dimension.

1.2.4 *Why Audio?*

Before we move on to the next chapter, this section provides an argumentation for our focus on the analysis of *audio recordings*, rather than symbolic music data.

The motivation to use audio data in a corpus analysis study can be quite simple: sometimes no symbolic format representation of a particular music corpus exists. Assembling a symbolic format dataset of music often requires tedious transcription work, especially in the case of music that is not part of the Western art music tradition ('classical music', in the broad, colloquial sense), whereas audio data is much more readily available. However, there are also more fundamental reasons why audio music data can be the format of choice for research.

Today's digital symbolic music formats are based on Western notation [36].[4] As many musicologists have argued over the last decades, there is a limit to the range of musical ideas and expressions Western music notation can represent, and to which extent. It evolved to suit the needs of composers and musicians within a set of traditions now denoted together as European art music, therefore it has never been intended to capture the particularities of non-Western musical styles, or even popular music (e.g., electric guitar solos or rap vocals) [79]. On a more fundamental level, most music notation isn't primarily intended to describe music at all, rather, it serves as a means to convey some of the necessary instructions for a performer to make a composition or arrangement audible. Even though Western notation depicts a more direct representation of the musical 'surface' than, e.g., guitar or lute *tablature*, by representing music in a relatively instrument-independent format, it nonetheless omits a wide range of very important features

4 and for the case of MIDI, keyboard music.

of a sound.[5] Most significantly, those features that are historically tied to performance rather than composition or arrangement: expressive timing and dynamics, timbre, tuning, ornamentation, etc. [36, 123].

Of course, most researchers will acknowledge this, but might be interested in a particular aspect of music that can be represented symbolically without too much loss of nuance. Harmony in a pop song, for example, can to some extent be represented as chords, or the main melody of a romantic concerto as a single monophonic sequence of notes. Even in these cases, however, such a perspective implies a conscious or unconscious choice not just to leave the performer out of the equation, but also the listener. When this is not acknowledged, these perspectives carry with them an implicit assumption that what can be notated contains or correspond to something the human perceptual systems 're-extracts' from the acoustic signal. This is not the consensus in music cognition. How many individual voices can a typical listener discern in polyphonic music? How is our listening affected by difference in salience between notes in a performance? How many different chords can a listener without formal music training tell apart? How does a jazz fan process the fast arpeggios of a virtuosic saxophone solo? Findings by Huron on the first question indicate strong limitations of the perceptual systems, with clear differences between expert musicians and other listeners [76]. Many similar, related questions remain unaddressed.

On the other hand, of course, audio representations have limitations of their own. First, an audio recording cannot capture every aspect of a musical performance in perfect detail either (e.g., spatial effects) or may distort it (e.g., dynamics) However, it does capture a lot; essentially, everything required to reproduce it again to a listener in the same way most music has been experienced by most listeners in the last 50 years: over some kind of speaker system (e.g., headphones). This still leaves the listener out of the equation, but at least it contains the necessary ingredients to apply most of what we know about the perception of music as part of any analysis.

5 In tablature, instructions are represented in relation to the instrument, e.g., by indicating which strings and frets are to be played on a guitar, rather than which notes.

This touches on a second issue: even if audio representations contain the necessary material, we may not have the analysis methods to extract meaningful perceptual-level information that we humans have access to effortlessly. Therefore, an important part of this thesis will be to assess this issue in depth, curate and develop a set of simple but plausible representations of harmony, melody and timbre, building on perceptually justified feature extraction and informed by a listener-centered perspective on music, as an alternative to those employed in the analysis of symbolic music data. And as progress is made on both our understanding of the perceptual representations that make up the human auditory system and on the technology that is available to model it, we can expect more such efforts to be made in the future.

In short, there is a very large potential in audio data for corpus analysis, for three reasons: there is simply much more audio data than symbolic data available for research, symbolic music representations have fundamental shortcomings in the way they represent much of today's music, and audio data allow for a more listener-centered approach.

1.2.5 *Conclusion*

Having reviewed a number of recent evolutions in music research, namely new musicology, music information retrieval, empirical musicology and music cognition, we can better situate our intended approach to corpus analysis in this complex interdisciplinary field.

First, we recognize that indeed, music, or a piece of music, is not some externally defined object, that can be understood in terms of absolute truths. However, an empirical approach to music research is possible that acknowledges these concerns.

We also concluded that much of the work done in computational musicology hasn't been striking the right balance between empiricism and the concerns of new musicology, e.g., Lerdahl and Jackendoff's influential but mostly descriptive models of music theory. Similarly, a lot of progress has been made in music information retrieval, but there,

too, the gap between the 'computability paradigm' and the critical perspectives that challenged it hasn't quite been bridged.

A better place to look for a connection between an empirical methodology and a critical perspective of music—as an abstract, intangible phenomenon—is in cognitive musicology, which balances these concerns by emphasizing the need of predictive models and an approach to music from the perspective of the listener.

We can now position the corpus analysis approaches we propose in this thesis between these observations. First, following Huron, we aspire to an approach to music research that is empirical, but avoids 'absolute' theories of 'absolute' music. Second, the corpus analysis approaches in this thesis will be quantitative, but not necessarily computational. Of course, computational models can be useful: tools from MIR or computational models of perception of cognition should be used if they help in providing a more relevant or cognitively valid representation of the data, but the 'computational' paradigm of ground truth reconstruction and its focus on performance should not be the main methodology. Third, our approach to corpus analysis will be informed by music perception and cognition, in that it assumes the perspective of the listener. The methodological consequences of these three positions will be discussed in chapter 3.

1.3 OUTLINE

1.3.1 *Structure*

This thesis is divided into three parts. The first part looks into related work and methodology. We give an overview of the audio description methods that have been proposed in the music information retrieval literature, concentrating on timbre, harmony and melody. We also give an overview of some of the applications for which they have been developed: music classification, structure analysis and content identification (chapter 2: Audio Description). In the next chapter, we review a selection of corpus analysis research, focusing on hypotheses, data, descriptors and statistical analysis methods. We also review, in a case

study, two important audio corpus analysis studies on the evolution of popular music [122, 175]. Based on the review and case study, we formulate a number of methodological guidelines for corpus analysis of music, and musical audio in particular (chapter 3: Audio Corpus Analysis).

The second part of the thesis presents contributions to support future audio corpus analysis research. The first chapter of this part centers on a corpus analysis of song sections in the Billboard dataset [26]. We present a first selection of relevant audio corpus analysis features for the analysis of harmony, melody and timbre. We also apply, for the first time, a feature analysis of audio data based on probabilistic graphical models (chapter 4: Chorus Analysis). The next chapter expands the available feature set for audio corpus analysis by presenting a new type of cognition-inspired melody and harmony descriptors. The third and final chapter of this part details the computational aspects of the descriptors presented in Chapter 5, and presents an implementation of the features for use in corpus analysis and content-based retrieval (chapter 6: Audio Bigrams). Both chapters connect the audio description contributions to the project goal of improving efficient music similarity measures for musical heritage collections.

The final part of the thesis addresses the last remaining research goal, bringing several contributions together in a computational analysis of hooks in popular music. The first of two chapters on this experiment presents Hooked, a game designed to collect data for the analysis and implemented as *Hooked!*[6] and *HookedOnMusic*[7] (chapter 7: Hooked). The second chapter of this part presents the analysis itself, using the descriptors proposed in Chapter 4 and 5, and a set of novel 'second-order' audio features. The results provide new insight into what makes music catchy (chapter 8: Hook Analysis).

6 `http://www.hookedgame.nl/`

7 `http://www.hookedonmusic.org.uk/`

1.3.2 *How to Read This Thesis*

Readers only interested in a subset of the work presented in this thesis may prefer to skip through to the chapters of their interest. For example, readers interested in an introduction to music information retrieval may find chapter 2 useful in itself.

Please refer to figure 1 for an impression of the relationships between the chapters. Chapters in the left column focus primarily on audio description: audio features, feature evaluation and implementations. These might be most interesting for readers with a background in music information retrieval or signal processing. Chapters in the middle column focus mostly on corpus analysis. These might appeal to readers mostly interested in methodology, and in our eventual findings. Finally, readers with absolutely no interest in technical details and statistics may find chapters 3, on methodology, and 7, on data collection, most welcoming.

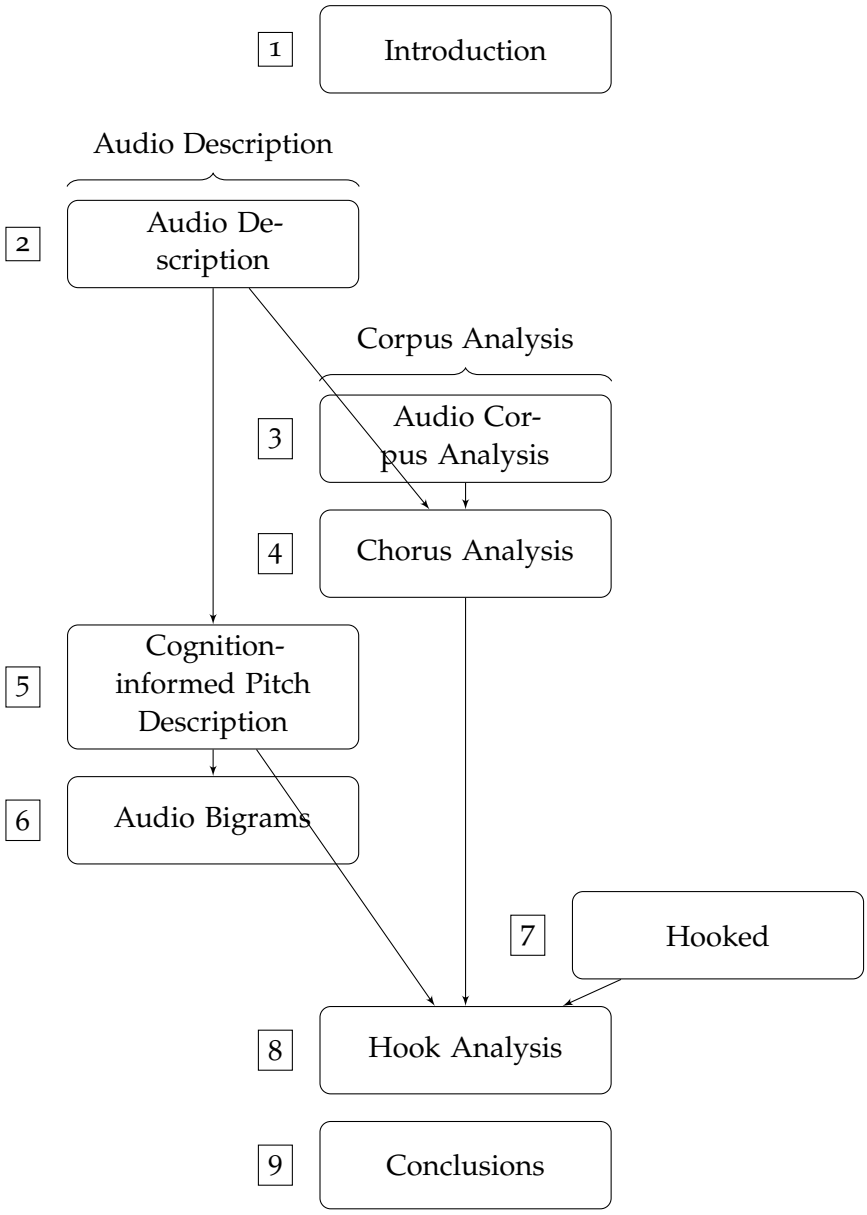


Figure 1.: Overview of chapter relations in this thesis.

2

AUDIO DESCRIPTION

Music information retrieval systems employ a multitude of techniques for the computational description of musical audio content. This chapter provides an introduction to this vast topic, by reviewing some of the most important audio descriptors (section 2.1) and a number of applications of these descriptors that are relevant to the rest of the work in this thesis (section 2.2).

2.1 AUDIO FEATURES

One conceptual framework that is typically used to write about and reflect on music descriptors distinguishes between low-level and high-level descriptions. Low-level features describe the properties of an audio sample on the level of the signal. High-level features correspond to the abstractions found in musical scores and natural language [179]. While low-level features tend to suit the language of machines and mathematics, high-level features are the ones that are used by humans (users of a music app, musicians, music scholars). And while machines are excellently equipped to make accurate measurements over a signal, humans, on the other hand, discuss and reason about music using a personal and highly 'enculturated' set of abstractions that changes over time and varies from individual to individual [5]. The discrepancy between signal-level and semantic-level music descriptions is complex: low-level descriptions may refer to not just signal-level, but also physical and sensory attributes of sound, and high-level representations can relate to formal, cognitive, or social aspects of it. The notion of mid-level features or descriptors is

sometimes also used to refer to an intermediate class of representations, e.g. in [107], where it is equated, roughly, to a level that aims to approximate perception.

The computational modeling challenges associated with this discrepancy are sometimes referred to as the 'semantic gap'. It has proven very challenging to teach a computer about these 'semantic' aspects of a piece of music. Modeling such high-level representations typically involves a trained classifier or probabilistic model learned from annotated data. Even advanced models, however, may not always yield reliable representations, and high-level representations often exhibit a significant trade-off between usefulness and reliability [179]. Questions have also been raised as to whether 'semantics', the elusive, high-level information some see as a holy grail of information retrieval, are present at all, in the audio representations used. The notions of 'semantics' and the 'semantic gap' may therefore be illusory, as Wiggins suggests in [202]. Nonetheless, the ambition to improve low- mid- and high-level representations to address increasingly high-level search problems continues to be a primary concern in music information retrieval.

The efforts put in by the audio content-based retrieval community have spawned an impressive collection of descriptors, low-, mid- and high-level. Many of these descriptions relate to one of the main 'dimensions' or parameters of music traditionally recognized in music theory: melody, harmony, rhythm and timbre. We will list a few of the most used and most relevant descriptors, with some explanation of how and why they may be used, in sections 2.1.1–2.1.4.

Features that do not directly fall into any of the categories along the low-level – high-level axis as introduced above include psycho-acoustic features and learned features. Psycho-acoustic features measure a specific psycho-acoustic attribute of a sound. They are often based on low-level signal measurements, but could be seen as high-level in that they approximate a human rating on a scale. Psycho-acoustic features are further discussed in section 2.1.5.

Learned features have come up more recently, out of the work done on feature learning for music content description. In feature learn-

ing applications, a content-based music description system might perform a classification of low-level music representations into high-level categories like genre, mood or a latent factor in a set of user behavior data, using advanced machine learning techniques. The classifier then yields, as an intermediate step in its pipeline, one or more novel, learned feature representations that are better suited to accomplish the original task (on the same or a larger dataset), or another one. Feature learning is discussed further in section 2.1.6.

2.1.1 *Basis Features*

The Fourier Transform

Many of the audio descriptors in this chapter are based on frequency information. To compute the amplitude of a signal at specific frequencies from an audio signal, the discrete signal $y(n)$ is converted to its frequency representation $Y(k)$ using the Fourier transform:

$$Y(k) = \sum_{n=0}^{N-1} y(n)\, e^{-2\pi i k n/N}, \quad k = 0 \ldots N-1 \tag{1}$$

yielding a single complex-valued spectrum for the entire time series $y(n)$.

Typically, the complex-valued $Y(k)$ will be represented in spherical coordinates, with *magnitude* $|Y(k)|$ and *phase* angle denoted $\phi(k)$:

$$Y(k) = |Y(k)|\, e^{-i\phi(k)} \iff \begin{array}{ll} |Y(k)| & \\ \phi(k) & = -i \ln(\frac{Y(k)}{|Y(k)|}). \end{array} \tag{2}$$

In many cases, only the magnitude of the Fourier transform is used.

To be able to look at the evolution of the frequency content, the Fourier transform is typically computed for an array of short overlapping windows in the signal, in a procedure called the *short term Fourier transform* (STFT). The result is a time series of windowed spectra, together forming the *spectrogram*.

$$Y(j,k) = \sum_{n=0}^{M-1} w(n)\, y(n + jH)\, e^{-2\pi i k n/N} \qquad \begin{array}{l} j = 0 \ldots \lfloor M/N \rfloor \\ k = 0 \ldots M-1 \end{array} \tag{3}$$

where M is the total length of the time series and N the length of the window w that is applied.

The frequencies to which k corresponds depend on the length of the window N and the sample rate of the original audio f_s (e.g., 44100 Hz):

$$f = \frac{k}{N} f_s \tag{4}$$

Furthermore, the Fourier transform of a real-valued signal $y(n)$ always yields a complex-valued spectrum $Y(k)$ for which the magnitudes $|Y(k)|$ are symmetric around $N/2$, and the phases are *antisymmetric* around $N/2$. As a result, all frequency information is contained in $k = 1 \ldots N/2$. The frequency f corresponding to $k = N/2$ is called $f_N = f_s/2$, the *Nyquist frequency*.

From this point on, we will abandon the standard notation for the discrete frequency analysis used above, in favor of the continuous conventions, i.e., we will express frequency representations in terms of f, rather than k, time as t instead of n. We will also assume frequency axes go up to f_N rather than f_s. Finally, throughout all of this thesis, we will keep using arguments for indexing; i.e., we write $Y(j,k)$ rather than $Y_{j,k}$; it will make formulas more readable.

Frequency Scales

The linear frequency scale f of the Fourier transform has some clear advantages. For example, a harmonic series of pure sinusoids will be represented as a series of equidistant peaks along the frequency axis, aiding a number of computations such as estimating the fundamental frequency of a complex tone. At other times however, a logarithmic division of the frequency axis may be more appropriate. Human perception of pitch follows *Weber's Law*. This law states that the smallest noticeable difference in some perceived quantity is, roughly, a constant percentage of the quantity in question, indicating a logarithmic sensitivity to the stimulus [111]. Similarly, the Western musical pitch scale follows a logarithmic function of frequency.

The constant-Q transform is similar to the Fourier transform, but follows a logarithmic division of the frequency axis [19]. Its name

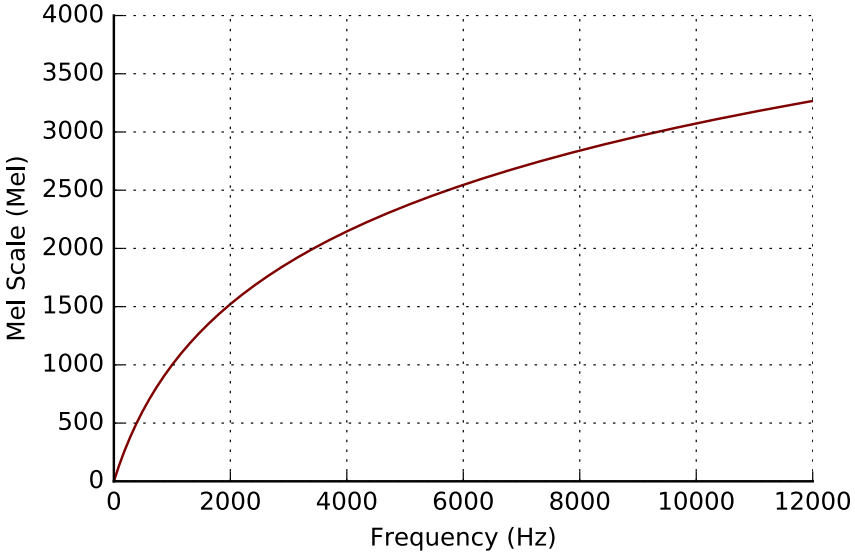


Figure 2.: The mel scale as a function of linear frequency.

refers to the constant relative width Q of the filters that would be used if the transform were to be implemented using a filter bank. The constant-Q transform and the constant-Q spectrogram are excellent basis features for pitch and harmony description and will be used as such in section 2.1.3. Some of the practical challenges in computing a constant-Q transform with useful resolutions are discussed in [137].

Experiments with human listeners have shown that, for a more accurate model of human pitch height judgements, Weber's law must be refined. Section 2.1.5 describes this in more detail, but in brief, the general observation is that the inner ear's representation of pitch is part linear with frequency, part logarithmic. One frequency scale that incorporates this is the mel scale (figure 2). It is roughly linear below 1000 Hz and logarithmic above 1000 Hz [111], but has no one formula. A commonly used formula is:

$$m(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{5}$$

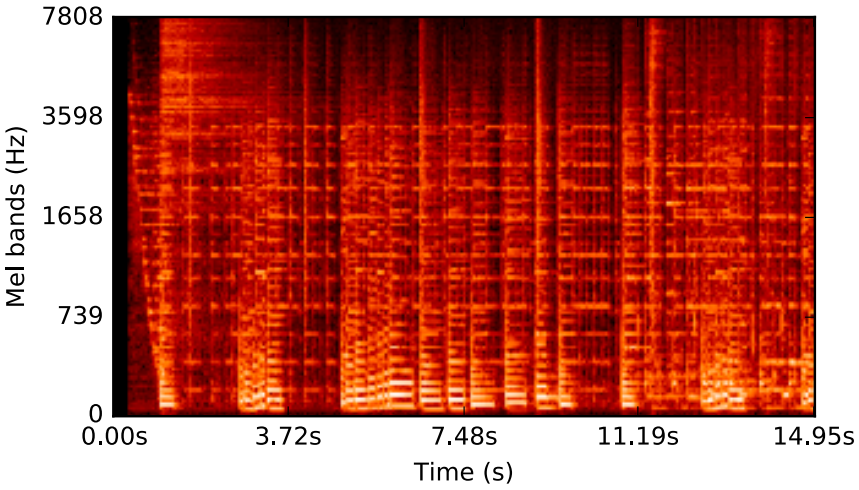


Figure 3.: Mel spectrogram. The song fragment shown begins with a piano playing a downward glissando.

This scale is the basis for much of the early work done in speech recognition, and a widely used basis feature for timbre description. The mel scale can be used to compute mel spectrograms, similar to the STFT-based linear spectrogram, but with a different frequency axis, as in figure 3.

2.1.2 *Timbre Description*

Timbre is a complex attribute of sound that is not easily defined. Timbre pertains to the 'tone color' or texture of a sound when pitch and loudness remain the same, and is crucial in our ability to recognize the differences between different instruments [30].

Timbre features, and particularly low-level timbre features, constitute a large number of the audio descriptors typically encountered in the music information retrieval literature. This may in part be explained by their success at predicting a particular notion of music similarity, mood, and musical genre. Some of these descriptors are

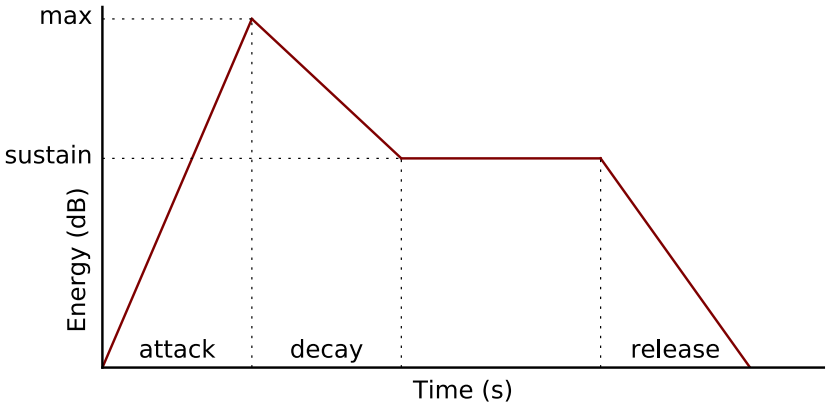


Figure 4.: Schematic of the attack, decay, sustain and release parameters of the temporal envelope associated with a note event.

now introduced, beginning with features that summarize the shape of the temporal and spectral envelopes of a sound.

Temporal Domain

In the temporal domain, the temporal envelope of a single note event is typically characterized by its attack time (the time it takes for its amplitude to reach an initial peak), its decay time after the initial amplitude peak, the sustain amplitude that is maintained until release of the note, and its release time after this point. Attack, decay, sustain and release are illustrated in figure 4. These descriptors are mostly useful for the description of single events, e.g. in the classification of instrument samples [154]. Another often-recurring time-domain is the zero crossing rate, the number of times the sign of an audio signal changes over a specified time window.

Frequency Domain

In the frequency domain, the spectral envelope is often parametrized by its first statistical moments, as if the spectrum were a statistical

distribution: the spectral centroid (the distribution mean), spectral spread (the distribution variance):

$$\text{centroid} = \sum_f f\, Y'(f) \tag{6}$$

$$\text{spread} = \sum_f (f - \text{centroid})^2\, Y'(f) \tag{7}$$

Spectral skewness and spectral kurtosis can be used too. These are the straightforward extensions of the centroid and the variance, again considering the amplitude spectrum as distribution:

$$\text{skewness} = \sum_f \frac{(f - \text{centroid})^3\, Y'(f)}{\text{spread}^3} \tag{8}$$

$$\text{kurtosis} = \sum_f \frac{(f - \text{centroid})^4\, Y'(f)}{\text{spread}^4} \tag{9}$$

where $Y'(f)$ is the magnitude spectrum, but normalized to sum to 1 [154]. Other spectral shape descriptors include the spectral roll-off point, marking the frequency below which 95% of the spectral energy is contained.

A very widely-used set of timbre descriptor are the MFCC features. Originating from the speech processing domain, mel frequency cepstrum coefficients describe the shape of the mel scale spectral envelope, by breaking it down into maximally de-correlated components: cosine-shaped basis functions referred to as *cepstral* components. Concretely: a mel scale amplitude spectrum $Y_m(m)$ is computed, after which the logarithm of the amplitudes is taken, and the coefficients of the components are obtained using a discrete cosine transformation (DCT) on the resulting envelope [127]:

$$\text{MFCC} = \text{DCT}(\log(Y_m(m))). \tag{10}$$

The DCT is a linear transformation in which a vector is expressed as a sum of cosines of different frequency $\{0, 2N, N, N/2, N/3, \ldots\}$. The coefficients of the summed cosines form the MFCC. Usually around 12 or 13 are used, of which the first is proportional to the sum of

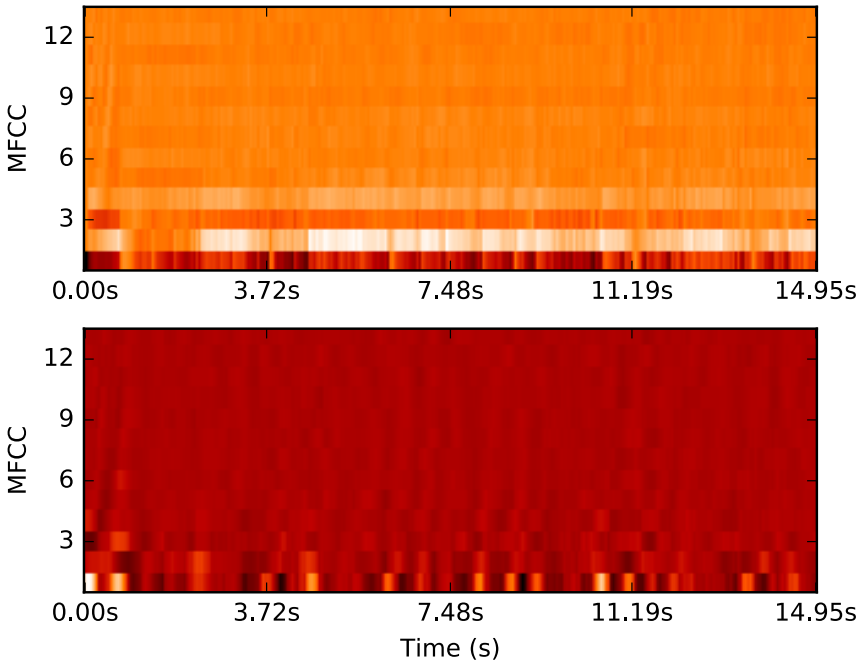


Figure 5.: MFCC and ΔMFCC features for a series of audio frames.

all amplitude bins, and therefore highly correlated to the spectral energy. When MFCC's are computed over a series of frames, it may be useful to compute so-called ΔMFCC and Δ^2MFCC features, the frame-wise differences of the MFCC and ΔMFCC, respectively. MFCC and ΔMFCC are shown in figure 5.

MFCC features have been used successfully as a basis for classification in a variety of tasks. As a result they have become the most used feature for the frame-based spectral (envelope) description, despite having been developed for non-musical applications first, and having seen some of its perceptual justifications refuted [6, 137].

Timbre features in general are widely used throughout MIR, perhaps because they have shown to work for a variety of tasks, perhaps because, as mostly low-level features, they are typically easy to compute and understand from an acoustics or signal processing perspec-

tive, whereas harmony and melody descriptors are more often rooted in music theory or music perception.

2.1.3 *Harmony Description*

Harmony, by definition, involves the simultaneous sounding of two or more pitches. Harmony description therefore relies on an accurate estimation of the pitch content in an audio segment. While the constant-Q transform provides a reasonable interface to this information, one particular alternative has turned out to be more practical: chroma features. We first introduce the notion of *pitch class* and the *pitch helix*, then discuss chroma.

Mathematical models of pitch

Both music theory and music perception describe a notion of *octave equivalence*. If a pitch is one octave above another, i.e., its fundamental frequency is two times the other's, the two are perceived as very similar. This effect is transitive: any two pitches that are spaced an integer number of octaves apart, will be perceived as similar. In the context of the standard equal-tempered scale, this establishes an equivalence relationship between each of the 12 pitches in an octave, and all of the pitches n octaves above and below it ($n \in \mathbb{Z}$). The resulting equivalence class is the *pitch class*.

Scholars since at least 1704 have proposed geometric models that integrate absolute pitch height and pitch class. Newton, in his book *Opticks*, first represented color on a color wheel, and linked this to pitch, as in figure 6 [139]. Donkin projects the pith axis on a spiral in a polar coordinate system, as in figure 7 [46]. Others have proposed representations that lie on the surface of a cylinder or cone, with pitch height along the central axis of the body and pitch class represented by the angle around this axis, resulting in a helix-shaped structure [100]. This enriched embedding has been adapted by Chew as the "spiral array" model of tonality, and used for visualisation purposes and to define harmonic distances [33].

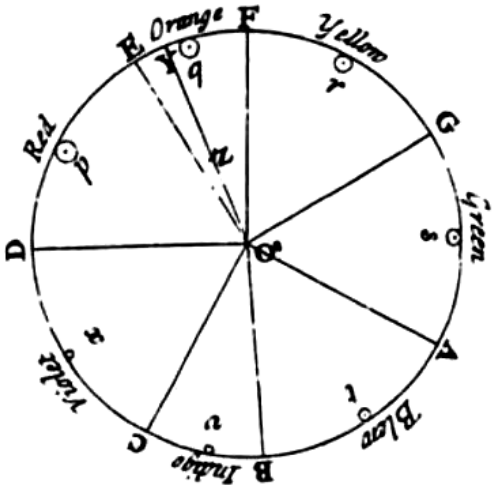


Figure 6.: Newton's circle representation of color and pitch. From [139].

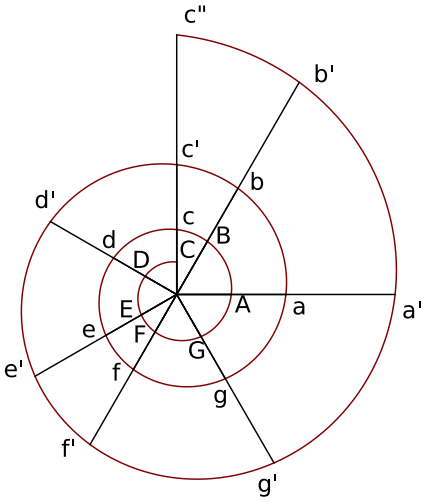


Figure 7.: Donkin's spiral representation of pitch. Adapted from [46].

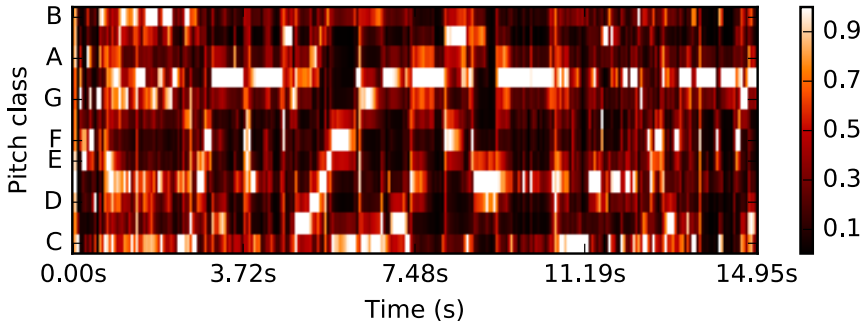


Figure 8.: A chroma time series.

Chroma Features

Chroma features are a representation of pitch class content that was first introduced by Fujimishima [54], as a feature for chord recognition. In its most basic form, chroma features or *pitch class profiles* (PCP), are a folded version of the constant-Q transform, in which the energy for all octave-equivalent pitches is summed together. Chroma features thus discard some of the absolute pitch height information, and only look at the pitch class dimension of the above spiral representation. Figure 8 shows a chroma time series computed this way.

Important advances in pitch description were made by considering the frequency-domain structure of complex pitches. The chroma features proposed by Gomez in 2006 do this by considering only prominent peaks in the spectrogram, summing together up to 8 harmonics per pitch and allowing for some deviation of the tuning frequency and the harmonic components, before folding the resulting harmonic pitch profile to a *harmonic pitch class profile* (HPCP) [55].

Wishing to reflect just pitch content and not the timbre information also present in the frequency spectrum, some pitch class representations take specific measures to achieve timbre invariance. Müller et al. discard timbre information by first computing the mel spectrum and setting the lowest 12-13 components of its DCT to zero before folding the amplitudes into pitch class bins [136]. This idea comes

from applications involving MFCC's, where it has been found that 12-13 coefficients are often enough to describe timbre in sufficient detail. Hence, keeping only the other coefficients might make for a good basis for timbre-invariant pitch class representation. HPCP features build in a similar invariance by "whitening" the spectrum prior to peak detection: a moving average of width m semitones is subtracted from the spectrum [55]. This approach is similar to Mullers, in the sense that both can be seen as a low-pass filtering operation on the spectral envelope.

Mid- and High-level Tonal Descriptors

The state-of-the-art in pitch and pitch class description involves a wealth of derivative features, most of which are built on the above representation. We review three families of mid- and high-level tonal descriptors.

Firstly, chroma features can be used to compute an estimate of the *tonal center* (e.g. tonic, in Western harmony), *mode* (e.g., major, minor), or *key* (tonic and mode) over a given time interval. This is typically done by estimating the correlation of the chroma features to a profile of pitch class occurrences. Commonly used profiles are the diatonic key profile and Krumhansl's empirically established templates [55,100]. Another, related harmonic descriptor is the *key strength* or *tonal/key clarity*, the confidence of the key estimate.

Harte and Sandler define the *tonal centroid* as the projection of the chroma vector in a complex geometric embedding. The 6-dimensional embedding can be seen as something in between the 12-dimensional chroma and the 2-dimensional circle of fifths. The tonal centroid feature can in turn be used to construct the Harmonic Change Detection Function (HCDF), a measure of harmonic change that is useful in (chord) segmentation of audio fragments [65]

In [56], Gomez and Herrera use a number of HPCP-based features to study the differences between Western and non-Western tonal music. Two relate to scale and tuning: the *tuning frequency* (deviation from 440 Hz) and equal-tempered deviation (average deviation from

an equal-tempered scale based on the tuning frequency), and are extracted from an HPCP with a resolution of 10 cents (0.1 semitones). They also compute the *diatonic strength*, the maximum correlation with a standard diatonic profile. Lastly, they include the *octave centroid*, the average pitch height computed before folding the pitch profile into an HPCP.

2.1.4 *Melody Extraction and Transcription*

Traditionally in music theory, *melody* is seen as the prominent, monophonic sequence of notes that characterizes a tune. Providing access to this sequence of notes, given a recording, has been one of the most elusive computational challenges in music information retrieval research. It is fair to say that, in the general case, complete reliable extraction and transcription of the main melody in a mix remains a challenging and unsolved problem, with state of the art accuracies of around 70% for the best systems– in just the extraction step [168]. The problem can be broken down into roughly three core issues. Firstly, separating components of a polyphonic mixture of complex sounds is very difficult. Mathematically, it is often an underdetermined problem, as the number of components is typically higher than the number of mixes (the latter usually being two in the case of stereo recordings). Humans solve this, to some extent, by employing a significant amount of top-down processing, i.e., using prior and contextual knowledge. Artificial systems can approximate these learned, contextual cues, but work on this is still catching up on the rapidly advancing technology coming out of the learning systems field. Secondly, estimating pitch from the separated stream is still a challenge in itself. To do this right, a system needs to decide when the melody is present and when it's not (voice detection), determine what note an octave is in, and pick out the main melody when notes overlap and sound simultaneously. Thirdly, identifying note boundaries remains a challenge for several instruments in which onsets (the beginnings of notes) and note transitions are blurred during performance (e.g., due to ornamentation) or production (e.g, reverb), including the singing voice. As the singing

voice often makes up the main melody, note segmentation on the main melody generally remains difficult.

We now review some pitch estimation systems based on three important strategies: the YIN algorithm (a time-domain approach) the Melodia algorithm (a frequency-domain approach), and an approach based on source-separation.

YIN

The YIN algorithm, proposed by de Cheveigne and Kawahara in 2001, is a pitch-estimation algorithm that operates in the time domain, i.e., no Fourier analysis is performed [40]. Instead, YIN works with a variant of the autocorrelation function (ACF) of the signal. The autocorrelation function scores the similarity of a signal with a time-delayed copy of itself. Formally:

$$\mathrm{ACF}(Y)(\tau) = \sum_t y(t)\, y(t-\tau) \tag{11}$$

The time delay τ is referred to as the *lag* and expressed in seconds. If a signal is exactly periodic with period T,

$$y(t) = y(t+T) \quad \forall T \tag{12}$$

the ACF will be maximal at $lag = T$:

$$\iff \mathrm{ACF}(Y)(T) = \sum_t y(t)\, y(t-T) = \sum_t y(t)^2. \tag{13}$$

but also at every multiple of T, including zero, as shown in figure 9. If the signal is noisy, but roughly periodic, the ACF will still reach a peak at every multiple of the period, so the method used to find T amongst these peaks needs to be considered carefully. Earlier systems based on the ACF were proposed by Hess and de Cheveigné [67]. Since, even with some measures in place, the autocorrelation method makes "too many errors for many applications", a cascade of error-reducing steps are added to the above procedure, to ensure that a sensible periodicity is found even if the signal isn't perfectly periodic, as is the case with

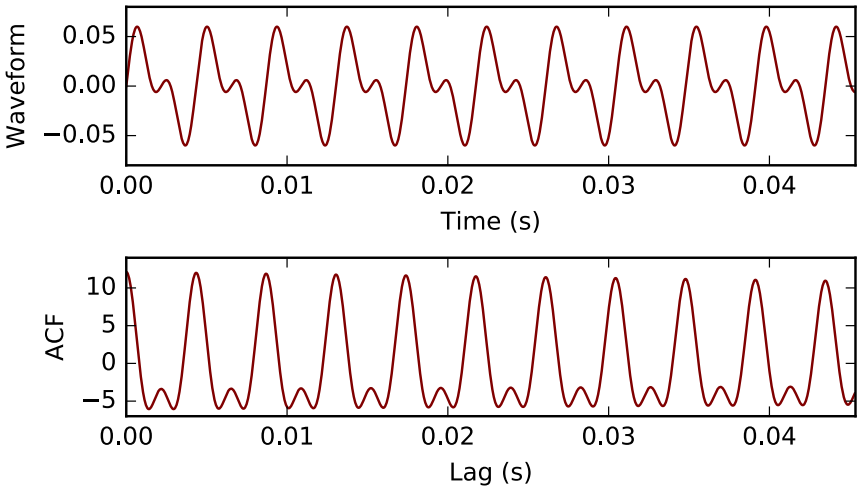


Figure 9.: A periodic signal and its autocorrelation function.

speech and musical pitch [40]. For example, instead of the ACF, the very similar autodifference function is found to increase performance:

$$\mathrm{ADF}(\tau) = \sum_t (y(t) - y(t-\tau))^2. \tag{14}$$

Finally, in another important modification to standard ACF-based approaches, over- and underestimates of T are avoided by not just looking for the global minimum of τ, but looking for the first value of $\mathrm{ADF}(\tau)$ that is substantially (e.g., 90%) lower than the average of ADF up to τ. This helps assess the significance of dips in the ADF near $\tau = 0$, where values are very low, as compared to dips for greater τ. Together, these measures constitute a robust procedure that has been shown to work for both speech and music. However, the signals on which YIN is typically used are largely monophonic, with only one, or one very dominant pitch present.

Melodia

Since the YIN algorithm, many have worked on main melody extraction from polyphonic signals: signals with several pitched sound sources present. Many of these systems start from a frequency representation of the signal, and assess, in different ways, the *salience* of all possible candidate pitches. One such system, which also performs well in comparative evaluations, is the Melodia system by Salamon.

Melodia pitch extraction is based on a high-resolution STFT, from which peaks are found. To get the best estimate of the exact location of these peaks, the *instantaneous frequency* is found by not just considering the magnitudes of the spectrum in each frame, but also interpolating between the phases of peaks in consecutive frames of the STFT [168]. Much like with HPCP (see 2.1.3), harmonic summation is then performed on the set of peaks (rounded to 10 cent bins) using a cosine weighting scheme. Extensive processing is also applied: peak candidates are grouped in time-varying melodic contours using a set of heuristics based on perceptual streaming cues. These candidate contours are then given a score based on their total salience and shape, and post-processed. The algorithm finally selects the set of contours that most likely constitutes the melody. In the latter step, the Melodia system also characterizes each contour as either voiced (sung by a human voice) or unvoiced, and decides in which frames no predominant melody is present at all.

The use of a contour representation has proven useful outside the core tasks of pitch estimation. In [169], Salamon and Rocha propose a number of mid- and high-level melody descriptors based on the contours extracted by Melodia, for a genre classification experiment. They include contour duration, mean pitch height, pitch height deviation and range, and presence and amount (width, frequency) of vibrato in each contour. Finally, each contour is also characterized as 1 of 15 contour types proposed by Adams, based on the order in which the highest, lowest, first and last pitch appear [2].

Data-driven and Source Separation-based Systems

A third group of melody extraction algorithms extract the melody by separating it from the rest of the mix. The simplest ones use a trained timbre model to describe each of two sources, one being the melody and another the accompaniment. These timbre models can be Gaussian mixture models (GMM's), in which each source is seen as a weighted sum of a finite set of multidimensional Gaussians, each describing a particular spectral shape, or hidden Markov models (HMM), a generalisation of GMM's. The models can be trained on source-separated ground truth data using expectation maximization [141]. Once the models have been used to separate the melody from the accompaniment, pitch estimation on the melody component is greatly simplified and a time- or frequency domain algorithm can be used. Advances were also made to apply newly developed machine-learning technology in the source separation step. In [180], Simpson et al. presented the results for a melody separation algorithm based on a deep convolutional neural network. The neural network is trained on spectrogram snippets of size 20×1025, yielding around a billion (10^9) parameters in total. The neural network approach improves on identifying main vocal melodies over a more traditional Non-negative Matrix Factorization-based approach. A purely data-driven model is presented by Poliner and Ellis. Skipping the separation step altogether, they use Support Vector Machines to classify STFT frames directly into melodic pitch categories [160]. The same authors have reviewed a number of other approaches in [161]. With the current trends in data-driven information retrieval methods trumping the performance of older, model-based approaches, more progress from this kind of methods can be expected in the near future.

2.1.5 *Psycho-acoustic Features*

Some of the features used in MIR relate to well-established psycho-acoustic qualities of sounds, e.g., loudness, sharpness and roughness. Each of these features quantises a perceptual attribute of sounds as

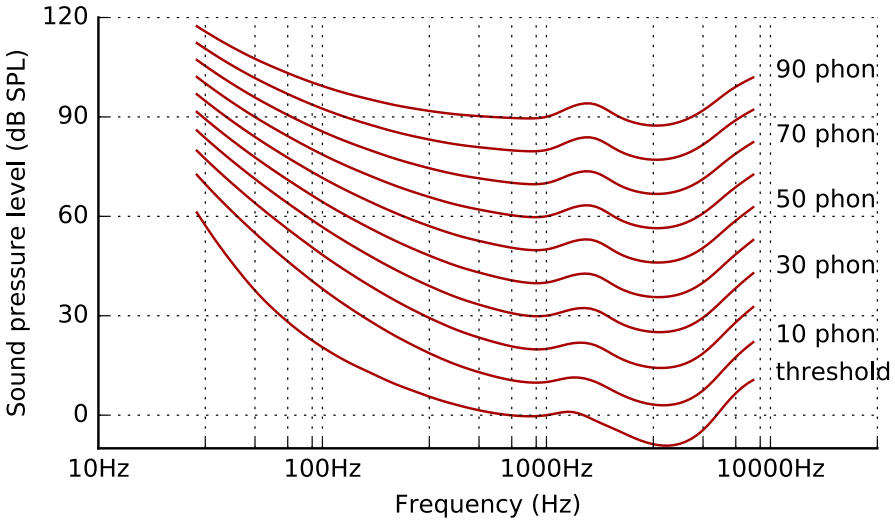


Figure 10.: Equal-loudness contours as specified by ISO standard 226.

rated by the participants of a listening experiment, or is based on (a model of) such measurements. The features take into account the non-linearities of the human auditory system.

The perceived *loudness* of a sound is determined by its intensity (as measured in dB) and its frequency content. At frequencies outside the 20–20,000 Hz range, sound is generally inaudible. But within the audible range, loudness varies with frequency as well. Equal-loudness contours specify this relation quantitatively. With these empirically established contours, shown in figure 10, a sound's loudness can be computed from its intensity at different frequencies. Two units of loudness exist: the *sone*, and *phon*. A single-frequency 1000 Hz sound at 40 dB has a loudness of 1 sone, and doubling a sound's perceived loudness doubles its value in sones. The phon is the basis of the ISO standard scale (shown in red in figure 10). A 60 dB SPL sound has a loudness of 60 phon. The phon scale is logarithmic: doubling a sound's perceived loudness adds 10 phon.

Loudness can also be computed for individual bands along the frequency spectrum. This yields an array of *specific loudness* values. A

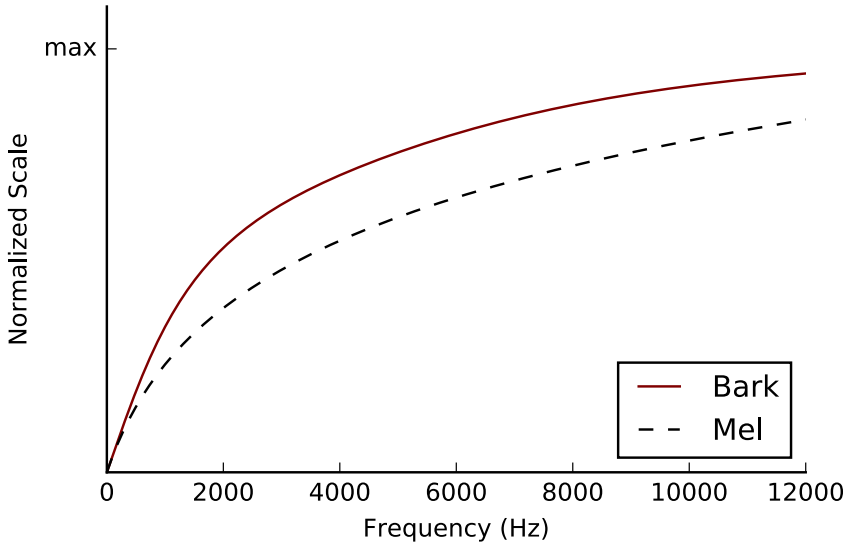


Figure 11.: The Bark and Mel frequency scales compared.

possible set of bands used for this purpose is the Bark scale. It is based on the mechanics of the inner ear. The arrangement of neurons along the inner ear's *basilar membrane* determines a *critical* bandwidth for every frequency. Within this band, masking occurs: the presence of one sound makes another more difficult to hear [111]. The Bark scale aims to take these elements of the frequency dimension's topology into account. Like the critical bandwidth and the somewhat simpler Mel scale, it is roughly linear at low frequencies (below 1000 Hz) and logarithmic at high frequencies (above 1000 Hz), as shown in figure 11.

Perceptual *sharpness* is a psycho-acoustic feature that is based on the Bark scale. While the above 'total' loudness integrates the specific loudness over all Bark bands, the sharpness feature measures the specific loudness distribution's centroid (i.e., center of mass). A sound for which the frequency content is more concentrated in the highest bands will have a high sharpness. Perceptual *roughness* is a result of

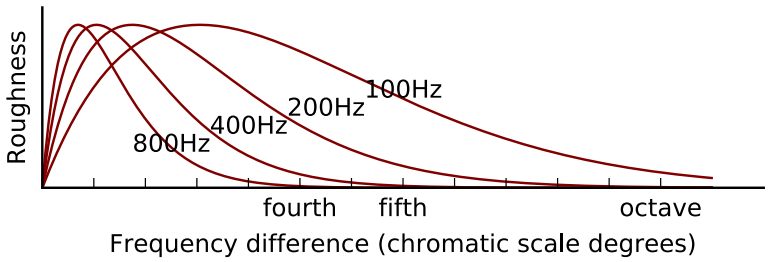


Figure 12.: Roughness as a function of frequency and frequency difference. Frequency difference is expressed in semitones.

the proximity of a sound's non-masked frequency components within the same critical bands. The roughness feature integrates the effect of these distances over the entire frequency spectrum. Somewhat simplified:

$$R(X) = \sum_{f_i} \sum_{f_j} w(f_i, \left|f_j - f_i\right|) X(f_i) X(f_j) \tag{15}$$

where w is a function of the first frequency f_i and its distance to the other frequency f_j over which is summed for every f_i. An example of different w for $f_i = 100, 200, 400, 600, 1000$ is shown in figure 12. Perceptual roughness is low for primarily harmonic, sinusoidal sounds and high for noisy and inharmonic sounds [159].

The above features not only correspond to empirically established attributes of sound, the attributes to which they correspond are also widely used in natural language description of sound. We argue that this makes them effectively high-level features. An analysis in which a trend for any of these descriptors is observed, can easily be translated back to domain language and natural language, making them an excellent instrument for computational research on musically motivated research questions.

2.1.6 *Learned Features*

As discussed in the beginning of this section, it can be useful to learn new representations entirely from data. This section reviews a number of techniques that can be used to do so, focusing on studies that don't just solve a particular task, but yield useful representations along the way. As with most trained systems, we can distinguish between supervised and unsupervised statistical learning.

Supervised learning generally requires a dependent variable that the learning system is trained to predict using a ground truth. New features can be constructed by taking the feature transformations that are learned in this process out of the trained system, to apply them somewhere else. For example, a multilayer neural network can be trained to predict labels for a set of labeled training data, so that, after it has been trained, one of the hidden layers can be used as more informative feature vector instead of the feature vector that was used as input, to address a different task. This is often referred to as transfer learning. In [193], a set of non-linear transformations of the STFT is learned by training a model to predict music listening statistics. The resulting representation is then successfully used in a number of different tasks and different datasets, showing that representations learned for one task can indeed be useful in a different context.

In unsupervised learning, the structure of an "unlabeled" training dataset itself is exploited to construct alternative representations. For example, a sparse auto-encoder is a neural network that learns a non-linear feature transformation. The new feature is taken out of a hidden layer (typically the only one) in this neural network, but instead of training the net to predict labels, it is trained to reconstruct its input while maintaining a sparsity constraint on the activations of the hidden layer. In [74], Humphrey et al. use non-linear semantic embedding (NLSE), a similar technique based on convolutional neural networks, to organize instrument samples in a low-dimensional space.

Dimensionality reduction and clustering techniques like PCA and K-means can be used as feature representations, too. In [35], Coates et al. showed that encoding a feature vector as an array of distances to

k cluster means can outperform other unsupervised feature learning techniques of similar complexity. This K-means approach has the advantage of having only one hyperparameter (k), and being efficient to train. Dieleman and Schrauwen applied a K-means representation to a tag prediction problem in [44].

As PCA and K-means are conceptually related, similar experiments have been done for PCA-based features. In [64], Hamel introduced principal mel spectrum components (PMSC). PMSC features are obtained using feature whitening and PCA on short arrays of mel spectrum frames.

Because of their probabilistic nature, statistical and neural network-based methods may appear to carry a suggestion of cognitive plausibility. Indeed: in a purely connectionist, statistical learning-centered perspective on cognition, learned features are a technology that may be, at the same time, optimal computational solutions to an engineering problem, and plausible cognitive models. However, not only is a purely connectionist view on music cognition generally disputed, current feature-learning methods are still far removed from realistic biological models of the brain, despite the quick successions of trends suggesting rapid progress.

2.2 APPLICATIONS OF AUDIO FEATURES

This section will provide a high-level overview of the music information retrieval field, focused on the problems, or tasks, researchers have addressed. We illustrate some of the most common practices in audio-based MIR research, to contextualize the origin of many of the features described in the previous section, and to give the necessary background for the critical discussion of these features in the next chapter. Most of the discussion, however, will be focused on those topics that are most relevant to this thesis.

In section 2.2.1, some of the most important work regarding music classification is reviewed, perhaps the core of 'classic' MIR, including the popular topics of genre and mood extraction from audio. In the next subsection, we review the most important methods in mu-

sic structure analysis, audio thumbnailing and chorus detection. This will be relevant to our work in Chapter 4. In section 2.2.3, the state-of-the-art in cover song detection and audio fingerprinting is reviewed. This will be relevant in Chapters 5 and 6.

2.2.1 *Audio Descriptors and Classification*

Audio classification tasks make up a large part of the most widely practiced research activities in music information retrieval, so they cannot be left out of a review of audio descriptor applications. But classification is also relevant because it can be seen as a form of high-level description, e.g., in terms of sociocultural information about the music. The resulting labels, then, are not properties of the music itself, but provide useful, user-level information for a variety of practical applications.

Genre classification

As one of the most widely researched topics in music information retrieval, genre classification deals with the automatic labeling of songs with genre tags. The appeal of this kind of information retrieval is easy enough to explain: originating in music sales and retail, genre tags provide a level of description that is useful in commercial contexts, and unlike many other descriptors used in MIR, genre and style labels are also widely used and understood by non-specialists [7].

The problem of genre classification allows for a very standard classification set-up: each document can be assigned one of a small set of class labels, and for each class a large set of examples can be found to train and evaluate classifiers on. Naturally, this involves some simplification, as genre description can be more or less detailed, and border cases are numerous.

To discuss all the audio features and classification algorithms that have been used in MIR would make for a very long and boring review. Most popular classification algorithms, like nearest neighbor classifiers, decision trees, support vector machines, neural networks

and random forests, have at some point or other been used to classify songs into genres [60].

One of the first and most influential studies to address music genre recognition (MGR) in depth is a series of experiments by Tzanetakis and Cook [188]. The study presents the first version of a now widely used dataset, *GTZAN*. A set of audio descriptors for MGR is proposed that includes tempo and rhythm features, timbre features (including MFCC), and some summary features computed from the pitch histogram. The classifiers that are studied are two classic density estimation models (a simple Gaussian model and Gaussian mixture models; GMM) and a non-parametric model (k nearest neighbours).

Since 2002, several improvements and variations were proposed that stick with the general approach of hand-crafting features and training a classical pattern matching classifier on the GZTAN ground truth, many of which were reviewed by Scaringella in 2006 [172] and Guaus in 2009 [60]. Notable additions to the above pipeline include features that build on improved models of the auditory systems, such as in the work by Panagakis [143], and the use of more powerful classification algorithms that have since emerged, such as support vector machines [115] and AdaBoost [11].

When, around 2010, feature learning techniques became widespread, MGR did not stay behind, and a variety of genre recognition systems were proposed that made use of technologies like learned sparse representations (e.g. [144]) and deep belief networks, a flavour of neural networks that are trained in a largely unsupervised manner [43].

Following these advances, classification accuracies reported in recent MGR studies have approached and exceeded the 90% mark on the GTZAN dataset [185]. It may be tempting to conclude that MRG is a solved problem, but as accuracies exceed even the GTZAN's theorized upper bounds due to inter-annotator disagreement, such claims taken with a grain of salt. This performance paradox has been explained by a combination of dataset issues (faults in GTZAN) and more fundamental issues around the usual approach to genre modeling, some of which will be discussed in section 3.3.3 [185].

The high performance numbers obtained in MGR may explain why feature learning researchers moved on to similar, but more difficult tasks, such as the more general 'tag prediction' task, with successes reported for convolutional neural network-based approaches and 'shallow' learning techniques such as k-means [44, 63].

Tag Prediction

In tag prediction experiments, a system is trained to predict manually assigned descriptive 'tags' for a dataset of songs. Contrary to MGR, tags can refer to any aspect of the music, including genre and style, but also instrumentation, language, topic of the lyrics, sentiment, geographic origin, era, mood, artist gender and form.

The rise of tag prediction and or 'social tag' prediction as a task can be traced back to the rise of the social web, where, on sites like *Last.fm*[1] and *MusicBrainz*[2], the enrichment of on line music data was crowd-sourced—by linking to social networks or through a Wiki-like platform.

Mood and Emotion Prediction

Another widely researched set of so-called top-level descriptors of music, are music mood and emotion. Music, in many parts of the world, is understood to fulfill a role as a 'tool' or medium for emotion regulation [71]. Application-oriented research efforts see emotion as an important practical attribute of music, that can be used in music search, recommendation, and in contexts like advertising and mood regulation apps.

Exactly how emotions are associated with specific pieces of music is a subject of debate. In music emotion literature, two mechanisms are typically distinguished. On the one hand, music can, to some extent, *express* emotions, through the intentions of the composer or performer. Emotional 'content' can then be perceived by the listener, though this perceived emotion isn't necessarily the same as what the artist intends

1 http://www.last.fm

2 http://www.musicbrainz.com

to express: the expressed emotions may not be perceived, or only selectively, while some of the emotional content perceived by the listener may not be intentionally communicated by the artist at all. On the other hand, there is the induced or felt emotion. This is the emotion that is induced in listening, and may be vastly different again from the emotion expressed by the artist or perceived to be expressed by the listener.

Much of the emotional value perceived in, and induced by music, is understood to originate externally to the music itself, in the listeners personal and cultural associations for example, or their social environment. This makes the task of automatic music emotion recognition (MER) from purely musical data difficult. A related task that circumvents the semantic subtleties of discussing emotion in music, is mood prediction. In this task, songs have been tagged with 'mood' labels, and a system is trained to reproduce these annotations. Whether mood prediction refers to a distinct task or merely a reformulation of perceived/induced emotion recognition, is beyond the scope of this discussion.

Many studies in MER work with an emotion space of just two dimensions: the valence-arousal plane, shown in figure 13 [206]. The idea is that all the emotions in this model can be situated along two principal axes. *Valence* relates to pleasure and distinguishes between positively and negatively experienced emotions. *Arousal* relates to energy or activation. Happiness, in this model, is a high-valence emotion characterized by a more or less average arousal. Anger is a low-valence, high-arousal emotion. Algorithms that use this space to model emotion simply need to predict both variables on a continuous scale, to describe a wide range of emotions, e.g. using regression models. Studies that have advanced this *dimensional* approach include work by Korhonen, Yang and Panda [96, 145, 207].

Other researchers have followed a *categorical* perspective on emotion. In this view, as well as in mood prediction, emotions and moods are treated much like tags: for each mood, a classifier or ensemble of classifiers is trained to predict its presence. The computational advantage of the valence-arousal models is that only two variables need to

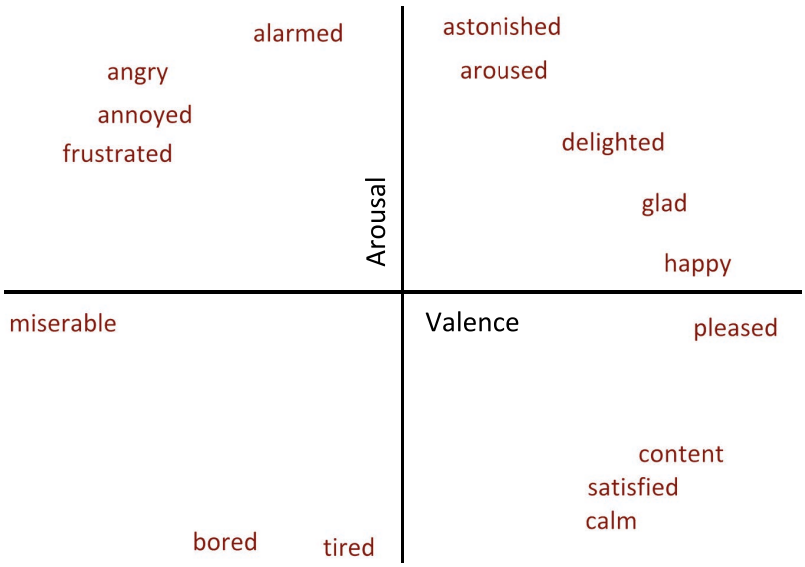


Figure 13.: The valence-arousal plane, a dimensional model of emotion (from [145]).

be modeled, compared to the very many (binary) variables typically involved in mood or tag prediction. The advantage of the tag prediction approach, however, is that the vocabulary doesn't have to be reduced to an agreed-upon space: mood tags might be include that do not seem to fit onto the plane at first sight (e.g., 'funny') or seem to collide (e.g., anger and fear) [91].

The first study to follow a categorical approach to mood and emotion recognition was done by Li and Ogihara and used 13 categories [109]. The audio mood classification task at MIREX uses a mood adjectives taxonomy based on 5 clusters. Others have used 4, 6, 8 and 18 clusters, to name just a few popular choices [206]. One music emotion taxonomy that allows for several 'resolutions' is the Geneva Emotional Music Scales (GEMS) model, a domain-specific model that was developed for music and allows for 9, 25 or 24 terms to be used. It was used in a study of induced emotion by Aljanaki in [4].

Summary

Throughout the many studies in MGR, MER and tag prediction, a number of recurring technologies and practices have come to fruition, often relying on a combination of audio features and classification schemes to reproduce high-level manual descriptions. Along the way, MGR and MER focus areas have carved out a practical and versatile approach to high-level music description—genre, tags, mood—that is powerful, but heavily reliant on machine learning.

2.2.2 *Structure Analysis*

In the MIR field of audio structure analysis, tools are developed to extract information from audio files on the level of structure or form. Commonly with the intent of using this information for further processing; in a few cases, as an end goal. For example, some applications of MIR benefit from prior segmentation of a recording, e.g. audio similarity computation [165]. Structure analysis as an end goal can be seen in a service like the Music Listening Station by Goto et al. [58].

Specific structural information retrieval tasks include structural segmentation, phrase segmentation, summarization, thumbnail extraction, chorus detection and full structure analysis. Structural segmentation refers to finding the boundaries of structural sections. In thumbnail extraction and music summarization, a stretch of audio is reduced to a one or more short subsections that are maximally representative of the recording [10]. Chorus detection refers to a similar task for popular music, in which the chorus of a song is located, to be used for indexing or as a representation in a browsing interface [57]. Structure analysis typically refers to structural segmentation, followed by a labeling of each segment with their structural function (e.g. verse or chorus in popular music, head in jazz, stanzas in folk music, exposition and bridge in classical forms, etc.) [38].

Two good overviews of structural analysis in the literature are provided by Dannenberg and Goto in [38] and by Paulus, Klapuri and Müller in [150]. The latter distinguishes between novelty-based, homogeneity-based and repetition-based approaches, echoing a distinction made first by Peeters in 2007 [155]. Peeters identified two general strategies: the state approach and the sequence approach. Most of the research follows one of these strategies; a few attempts have been made to combine both approaches. The most important contributions will now be explained, following Peeters' distinction.

State-based Structure Analysis

In the *state* approach on structure analysis, a song is interpreted as a succession of observable states, which can be mapped to structurally meaningful sections or 'parts'. A state spans a contiguous set of times during which some acoustical properties of a song are more or less constant. This is said to hold for popular music, in which the 'musical background' often remains the same throughout a structural section. The state approach is applied mostly in combination with timbre features, such as MFCC, since they tend to correlate with instrumentation [150].

State representations can be obtained in various ways. The *novelty approach*, for example, detects transitions by looking for peaks in a novelty function. A naive novelty function can be constructed by correlating a feature time series with a length N novelty kernel such as

$$z(n) = sign(n) \cdot \Phi(n), \quad n = -N/2 \ldots N/2 \tag{16}$$

where Φ is some symmetric Gaussian. Other approaches apply HMM or similar methods to group frames of features into states, often using two (or more) techniques sequentially. Clustering the obtained states finally allows for the mapping of the observed time spans to more or less meaningful musical parts.

Responding to the field-wide growing appeal of data-driven methods, Ullrich et al. recently proposed a relatively simple method using convolutional neural networks (CNN) [190]. A CNN is trained to predict the presence of a boundary given a region of frames of a basis feature, with good results. The network does not keep any history, so the cues it uses can be assumed to relate to novelty rather than repetition, making this method effectively a state-based one.

Sequence-based Structure Analysis

The sequence approach relies on repetitions of sequences of features to infer structure. The frames in a sequence do not need to show any similarity amongst themselves, as long as the sequence as a whole can be matched to a repetition of the sequence somewhere else in the song. Most repetition finding approaches work on the *self-similarity matrix* (SSM) of the song.

The self-similarity matrix is an essential tool of many structure analysis algorithms. For music feature representations that are computed over short frames, the *self-similarity matrix* (SSM) is a matrix representation of the similarity of each frame to every other frame:

$$\text{SSM}(i, j) = s(X(i), X(j)) \tag{17}$$

where s is a function measuring similarity between two frames of audio. SSM matrices reveal state structure as homogeneous blocks, while

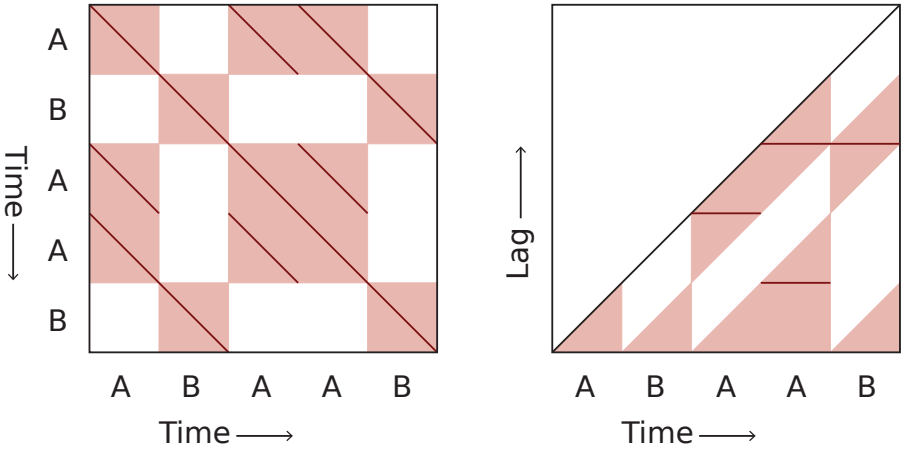


Figure 14.: An idealized SSM and corresponding time-lag matrix. Darker regions denote higher similarity. The state structure can be seen as blocks while the sequence patterns are visible as 'stripes'. Adapted from [150].

repeated sequences emerge as diagonal 'stripes' as shown in figure 14 (left). Sequence-based approaches using SSM focus on these stripes, and use chroma as its basis feature $X(t)$.

Variants of this approach include the use of a *self-distance matrix* (SDM), which contains the distance between two frames d rather than the similarity, or the *recurrence plot* (RP), in which a short history of frames is used to assess similarity. Common distance functions are simple Euclidean and cosine distances.

The SSM's *time* × *time* representation of self-similarity can be converted to a triangular *time* × *lag* matrix (shown in figure 14 on the right) through appropriate 'folding', i.e. $(i, j) \mapsto (i, i - j)$, in which the lag $i - j$ is the time difference between frame i and j. *Time* × *lag* matrices conveniently show repetitions as vertical stripes. Some methods make use of a beat-synchronous SSM to account for tempo variations within the song, or the transposition invariant SSM introduced by Clausen and Müller [34].

Finding repeated sequences in the SSM is not as straightforward as it may seem and greatly benefits from the post-processing of the SSM after it has been computed. A moving average filter can be used to smoothen the matrix along the columns or diagonals as in [9, 57], as well as erosion and dilation (two grey-scale image processing operations, often combined to remove short interruptions in a uniform sequence) [50,114]. A high-pass filter can in turn be used to emphasize details in the opposite (lag) direction. The result is then typically converted to a binary SSM by comparing to a constant or relative threshold.

Finally, a set of repetitions is extracted from the SSM, each identified by a start, end and lag time. Two strategies can be observed. Goto's RefraiD algorithm [57] extracts repetitions by first looking for those lag times corresponding to the lowest distances. Along these columns or rows, it then stores all appropriate length time intervals in which the binary distance value is zero. Unless the feature frames are beat-synchronous, this method doesn't allow any deviations in timing or tempo. Another method proposed by Chai [32] uses dynamic time warping (DTW, see section 2.2.3) for the alignment of sequences to account for local tempo variations. Computationally, this is not very efficient since many different sub-matrices need to be matched (one for each pair of candidate sequences). Variations based on dynamic programming were used by Dannenberg and Hu [39], Paulus and Klapuri [147, 149] and recently by Müller, Grosche and Jiang [137].

Combining State and Sequence Representations

A number of recent methods have combined steps from these state and sequence approaches to advance the state of the art in structural segmentation accuracy. In [85], a simple combination method is proposed for the fusion of two independently obtained sets of candidate boundaries.

A more sophisticated method was proposed by Serrà in [175]. In this method, a newly proposed variant of the lag matrix is filtered along the time axis with a step function kernel, with the aim of de-

tecting the start and end points of individual repeated segments. The resulting 'structure features' matrix is then summed over the lag axis to obtain a novelty curve quite like the one typically used in state-based segmentation methods. The method is reported to work for both timbre and pitch features, and performed better than its competitors in the MIREX 2012 audio structure analysis track. Peeters et al. developed this idea further by combining it with a prior analysis following Goto's approach described above [156].

Another method proposed by McFee in 2014 integrates state- and sequence-based methods using a graph representation of the audio, rather than an SSM [125]. Each frame of the base feature X is represented as a node in a large graph, and edges between nodes are weighted by the similarity between the frames. A technique called spectral clustering is then applied to obtain a hierarchy of section boundaries, from which the best set of boundaries can be obtained in a supervised way, by letting a user or 'oracle' select the most appropriate level of segmentation.

Thumbnailing and Chorus Detection

As implied in the introduction to this section, summarization and audio thumbnailing are practically the same task. The term thumbnailing was introduced by Tzanetakis in [189]. Chorus detection is de facto a form of thumbnailing, specific to popular music, in which one wishes to locate the chorus of a song.

Definitions of chorus often make sure to include that a chorus is *prominent* and/or *catchy*, though this is rarely explained or formalized. Both thumbnail and chorus are essentially reduced to the *most often-repeated segment*. Just like in much of structure analysis, most research is therefore devoted to finding these repeated segments, combined with minor heuristics limiting the candidates. More advanced approaches include the system by Goto [57] and Eronen [50].

Any of the repetition-detection methods discussed above can in principle be used, and many of them come from papers on summarization and chorus detection. We conclude this section with an overview

of the heuristics that are used to select the most representative repetition.

After the obtained repetitions are 'cleaned' (using some heuristics for boundary refinement and dealing with overlap), they may be clustered to obtain meaningful groups, each corresponding to a part of the song, like in full structure analysis. Transitivity may be exploited here: if b is repetition of a and c is a repetition of b, then c should be a repetition of a. The grouping task, especially important in full structure analysis, is not trivial either. Cost functions and fitness measures have been proposed to rate the amount to which a proposed structure explains the observed patterns [137, 149, 155].

Scoring functions for the assessment of thumbnail and chorus candidates have been proposed as well [50, 57], introducing a variety of heuristics. The RefraiD system by Goto [57] makes use of a scoring function that favors segments c occurring at the end of a long repeated chunk abc and segments cc that consistently feature an internal repetition. Eronen [50] favors segments that occur near $1/4$ of the song and reoccur near $3/4$ as well as segments with higher energy. In most cases, heuristics are only used to limit the candidates from which the most frequent segment is picked. For example, by considering only the first half of the song or discarding all segments shorter than 4 bars.

Regarding chorus detection, it is clear that existing strategies in chorus detection only attempt to locate refrains in a pragmatic way, and do not aim to model what choruses are and what makes them distinct. This problem is addressed as part of this thesis, and detailed in chapter 4.

2.2.3 *Audio Fingerprinting and Cover Song Detection*

Audio fingerprinting and cover song detection systems both deal with the automatic identification of music recordings.

Robust, large-scale audio fingerprinting was one of the first problem in music information retrieval to be convincingly solved, and developed into a successful industry product. Effective audio fingerprinting algorithms like the ones developed by Haitsma and Kalker at

Philips [62] and Wang and Smith at Shazam [199] can reliably identify a single exact music fragment in a collection of millions of songs. This is useful as a service: the Shazam algorithm stands as a very popular app, and was even available before smartphone apps, as a phone service. But the technology can also used for content identification of on line radio, and on social networking sites and streaming services like Youtube[3] and Soundcloud[4]. Last but not least, fingerprinting can also be used to manage large collections and archives, e.g. for duplicate detection.

In cover song detection, or (cover) version identification, a system is charged with the task of matching a recording to other known versions of the same musical work, generally interpreted by other artists. This can be useful in a similar set of applications, most notably content identification and duplicate detection, but also plagiarism detection and music recommendation [174].

Audio Fingerprinting

Audio fingerprinting, at its core, involves the reduction of a large audio object to a compact, representative digest. Given an unlabeled fragment, fingerprinting systems extract this fingerprint and match it to a large reference database. State-of-the-art algorithms for audio fingerprinting produce fingerprints with a high degree of robustness to noise, compression and interference of multiple signals, and perform matching of fingerprints very efficiently [31, 59].

The first widely successful fingerprinting technique was proposed by Wang and Smith, the so-called *landmark-based* approach [199]. Like most fingerprinting systems, Wang's system includes an extraction and a matching component. In the extraction component, a piece of audio is first converted to a spectrogram representation using the STFT, and the most prominent peaks in the spectrogram are detected. Peaks are then paired based on proximity. Pairs of peaks are called landmarks, and can be fully described by 4 parameters: a time stamp,

3 `http://www.youtube.com`
4 `http://www.soundcloud.com`

the frequencies of both peaks, and the time interval between them. In a last step, the two peaks frequencies and the time interval are combined into a hash code for efficient look-up.

The reference database is constructed by storing all hashes for a collection of songs into an index, where each hash points to the landmark start time and a song ID.

In the matching stage, when a query is passed to the system, its landmarks and corresponding hashes are computed as described above. Any matching landmarks from other songs are then retrieved from the reference database, with their corresponding start time and song ID. Note that this can be done in constant time. In the last step, the system determines for which landmarks the start times are consistent between query and candidate, and the song with most consistently matching landmarks is returned as the result.

Haitsma and Kalker's approach, developed around the same time, also relies on the indexing of structures in the spectrogram, but doesn't involve the detection of peaks at a high frequency resolution [62]. Instead, the spectrogram's energy is computed in 33 non-overlapping bands, and the resulting time series is differentiated across both time and frequency. The resulting 'delta spectrogram' is then binarized by considering only the *sign* of its values. In the extraction step, strings of 32 bits are extracted from this representation, and stored as sub-fingerprints, much like the landmarks in Wang's approach. The matching step follows a similar logic as well.

Alignment-based Cover Detection

Despite the conceptual similarity between audio fingerprinting and cover song detection, state-of-the-art audio fingerprinting and cover detection algorithms share very little of their methodology. This is largely due to the many invariances that need to be built into any cover detection technique: any two performances of a song may vary in just about every aspect of their musical material, and still be regarded cover songs [177]. A good identification system should there-

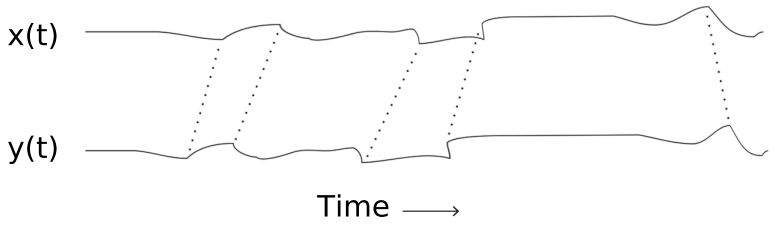


Figure 15.: Diagram showing the alignment of two time series performed in dynamic time warping. Adapted from [135].

fore be invariant to changes in tempo, key, structure, lyrics and instrumentation as well as, to some extent, melody, harmony and rhythm.

Cover detection systems have addressed this challenge in several ways. For example, tempo invariance, by representing songs as beat-aligned time series before matching [49], or key invariance, by performing a search using multiple, transposed, queries [102]. In [176], Serrà et al. propose a method based on the alignment of pairs of chroma time series that is not the first in its kind, but successful due to the incorporation of several novel invariance measures. Most notably, key invariance is achieved by first comparing the pitch histogram for each pair of songs, and transposing them to a common key. Tempo invariance is achieved using dynamic programming local alignment (DPLA) in [176], a form of locally constrained dynamic time warping.

Dynamic time warping (DTW) is an algorithm for the pairwise alignment of time series. In DTW, two time series are 'warped' in a non-linear fashion, to match each other in a maximum number of similar positions, as shown in figure 15. A score is assigned to each of several possible configurations, based on the similarity of matching frames and the number of skipped frames in each time series [135].

In [178], $Q_{\max}$ is introduced instead, a cross-recurrence measure defined on the cross-recurrence plot (CRP). The latter stems from non-linear systems analysis, and can be seen as a variant of the similarity matrix (SM) between two time series, where, like in the SSM (Section 2.2.2),

$$\mathrm{SM}(i, j) = s(X(i), Y(j)) \tag{18}$$

with s a similarity measure. The cross-recurrence plot differs from the standard similarity matrix by incorporating some 'history' of length m, the *embedding dimension*. This history is encoded in so-called delay coordinates:

$$\begin{aligned} \vec{X}_m(i) &= X(i-m:i) \\ \vec{Y}_m(j) &= Y(j-m:j) \end{aligned} \tag{19}$$

where : denotes a range. The CRP is given by:

$$\mathrm{CRP}(i,j) = s(\vec{X}_m(i), \vec{Y}_m(j)). \tag{20}$$

The cross-recurrence measure $Q_{\max}$ essentially measures how long the longest alignable segments are. Algorithms based on $Q_{\max}$ have performed best on the MIREX audio cover songs identification task since 2007 [178].

Scalable Cover Detection and Soft Audio Fingerprinting

While similar in concept, it is now clear that audio fingerprinting systems and cover song detection systems, as described above, are vastly different in their approach. This section will look at the commonalities to both strands of research, and lay out the prior art that combines ideas from both fields to address a common underlying problem. This work will be expanded on in chapters 5 and 6.

As stated at the beginning of this section, the common underlying problem between audio fingerprinting and cover song detection is the automatic content-based identification of music documents. Assuming a trade-off between efficiency and accuracy, we could say audio fingerprinting is an efficient solution to this problem, but not a very robust one: solutions are robust to several kinds of distortions to a query, but not to the wide variety of deliberate changes that cover detection systems take into account. They are unable to identify covers, live renditions, hummed versions, or other variations of a piece.

Conversely, the cover song detection systems reviewed above handle these modifications, but do so in a much less efficient manner. All of the cover detection systems reviewed in the previous section rely on some kind of alignment to assess the similarity for every pair

of songs. Since each query is linear in the size of the dataset N (N alignments are needed), and each alignment polynomial in the song lengths m and n, alignment-based algorithms are not a good solution for large-scale cover song retrieval [75].

Several efforts were made to adapt the concept of fingerprinting to such use cases, which require invariance to intentional, performance-related changes to the song. Relevant work includes a growing number of studies on 'scalable' cover song detection, pitch- and tempo-invariant fingerprinting, including sample identification, and some of the work done on 'query by humming' (i.e. identifying a song from a hummed or sung melody, generally performed by an amateur singer using a dedicated retrieval system).

In this thesis we refer to all of these tasks together as *soft audio fingerprinting* systems. The defining distinction between soft audio fingerprinting and other kinds of document retrieval, is the fixed size of the representations, which enables the use of an index to store them—guaranteeing the constant-time look-up.

Some of these soft audio fingerprinting systems follow Wang's landmark-based strategy, but build in some invariance. Audio fingerprinting systems targeting invariance to pitch-shifting and/or time-stretching include [51] by Fenet et al. and *Panaka*, by Six et al. [181]. Van Balen et al. [192] and Dittmar et al. [45] present automatic approaches for the identification of samples used in electronic music production. In each of these studies, the basis feature from which the peaks are combined into landmarks, is the constant-Q spectrogram, rather than the spectrogram.

Another landmark-based retrieval system is the large-scale cover song identification system proposed by Bertin-Mahieux et al. in [12]. Here, landmarks are extracted from pitch class profiles or chroma features (Section 2.1.3). As in fingerprinting, matching landmarks (here: 'jumpcodes') are retrieved with their song IDs. The study reports a mean average precision of about 0.03% and a recall of 9.6% on the top 1 percent of retrieved candidates in a large dataset: promising, but nowhere near the performance of alignment-based algorithms in their respective use case.

A more novel audio indexing feature, the *intervalgram*, is proposed by Walters [198]. It is essentially a two-dimensional histogram of local pitch intervals at various time scales, designed for hashing using wavelet decomposition. Another novel approach by Bertin-Mahieux uses a 2D Fourier Transform of beat-aligned chroma features to obtain a compact representation that is invariant to pitch shifting and time stretching [13]. This method was adapted by Humphrey et al. to include a feature learning component for more robustness to common variations [75]. The latter currently performs best in terms of large-scale cover song retrieval precision, though, with a mean average precision of 13.4%, still not close to the alignment-based state of the art.

The progress on some related tasks, such as query by humming, has been better. There is little information about the exact workings of commercial services such as Soundhound's MIDOMI[5], but they work well enough for commercial use. However, they are generally understood to rely on matching (alignment or otherwise) of simplified contours of melodies sung and labeled by volunteers, rather than matching hummed melodies with a song's original audio, which remains an unsolved problem.

2.3 SUMMARY

We have reviewed, in section 2.1, a selection of topics and state-of-the-art methods for audio description. We have focused on timbre description, harmony and melody description, psycho-acoustic features and learned audio descriptors. Several of these reviewed audio features and applications will be applied and improved upon in part ii of this thesis.

In section 2.2, we have reviewed a selection of the music information retrieval applications for which these features were developed. Here, we focused on classification and labeling tasks (genre, tags, mood and emotion), structure analysis, and music content identification (audio fingerprinting and cover song detection). In the last category, we have

5 http://www.midomi.com/

defined *soft audio fingerprinting* as the umbrella task of scalable music content identification, including efficient cover song detection, pitch- and tempo-invariant fingerprinting, sample identification, and query-by-humming. The task of soft audio fingerprinting is both an important open issue in MIR, and closely related to the project goal of modeling document similarity in musical heritage collections. Therefore, it will also be given more attention in the following chapters.

3

AUDIO CORPUS ANALYSIS

This chapter will provide more context for the research in this thesis, though not in terms of its technological context, like the previous chapter, but in terms of its methodology: *audio corpus analysis*. A large majority of the studies in computational music analysis center around transcription, classification, recommendation and retrieval, effectively limiting themselves to the reconstruction of a ground truth and rarely leveraging the power of computation to mine music collections for novel musical insights. We discuss the challenges and pitfalls in applying MIR's technology in the pursuit of a better understanding of music. At the end of this section, a selection of desiderata for dedicated audio corpus analysis technology is given.

3.1 AUDIO CORPUS ANALYSIS

3.1.1 *Corpus Analysis*

In this thesis, we define corpus analysis as: any analysis of a collection of musical works in which the primary goal is to gain insight into the music itself. Consider, as an example, Huron's study of melodic arcs in Western folk song [77]. In this study, Huron used the *Humdrum* toolkit and a corpus of 6251 folk songs from the Essen Folksong Collection to show a tendency towards arch-shaped melodic contours. Particularly, he demonstrated that, of nine simple contour-types, a convex shape was most common, and that there is a significant tendency for ascending and descending phrases to be linked together in

pairs. As we argued in chapter 1, corpus studies like this form part of 'empirical musicology', distinguish themselves from a large body of other computational music research by answering a musicological question and aiming at new musical insights.

3.1.2 *Audio Corpus Analysis*

Audio corpus analysis can now be defined simply as corpus analysis on audio data, as opposed to symbolic data (such as scores) or manual annotations. In practice, music data may not always come in an unambiguously unimodal form, but it is safe to say that there is a striking prevalence of symbolic datasets in corpus analysis. A review in the next section will show that audio is used only in a minority of the studies, despite its potential for corpus analysis as argued in section 1.2.4: despite its availability and despite being, by far, the most widely used and researched form of information in the music computing community.

Additionally to the arguments presented in chapter 1, recent audio corpus analysis results have gathered a wide interest in the press, notably, since work on this thesis began, Serrà and Mauch's studies of the evolution of popular music [122, 175] (see section 3.4) and the first results from the *Hooked!* game (see Chapter 7). Meanwhile at a more general level, too, the promise of leveraging bigger datasets in science and the humanities, has drummed up popular interest in data-rich research across the sciences and on the intersection of disciplines, see e.g., Leroi in the New York Times [106]. There is a vast, unexplored potential in using MIR technologies to answer questions about the increasingly abundant resource that are audio collections.

Now that the notion of corpus analysis and audio corpus analysis have been outlined, a selection of prior research will be critically reviewed in the next sections, focused on linking the first, the most influential, and most recent contributions.

3.2 REVIEW: CORPUS ANALYSIS IN MUSIC RESEARCH

We structure this review by distinguishing between three data formats: manual annotations, symbolic data, and audio data, where the category 'audio data' includes any recording of a piece of music, and 'symbolic data' roughly corresponds to machine-readable music notation (including scores, chord labels, digitized tablature and MIDI). Manual annotations refer to any set of manually assigned labels. The distinction between audio and symbolic, symbolic and encoding, may be somewhat artificial at times, but it is useful enough to guide us through the history of music corpus research. Finally, any overview like this is necessarily incomplete. The work that is included is selected to represent a variety of subdomains of music research. We explicitly exclude work with an important retrieval (e.g., search or classification) component, even if much of it has been very important to corpus analysis, e.g., classification of folk songs into tune families. We also haven't included much work from performance studies, a discipline that frequently employs music corpora as well.

3.2.1 *Corpus Analysis Based on Manual Annotations*

One pioneer who envisioned what can be considered the first 'data-driven' approach to the study of music culture, was Alan Lomax. Working as as ethnographer in the USA during the 1930's and 40's, and in England and Europe during the 1950's, Lomax made field-recordings of folk singers and musicians. Later, in the 1960's, he contributed to the foundations of ethnomusicology and performance studies with the *Cantometrics* methodology, constructed by Lomax and Victor Grauer in 1963. Cantometrics is a *coding* system in which recordings of sung performances from around the world are assigned scores on the basis of their stylistic properties and the social context in which the music is performed [112]. These annotations include how many singers participate in a performance, the melodic complexity, and how much vocal embellishments were used, among many other things. Lomax' intention was to correlate these ratings to other aspects of culture.

In one study, for instance, data from over 4000 songs out of the Cantometrics program were used to show that the prevalent performance style of a culture reflects the 'degree and kind of group integration that is appropriate and necessary to the culture's adaptive structure' [113]. Since the study of musical cultures in the world evolved, however, Lomax' perspectives on the evolutionary hierarchy of human cultures have been criticized, as well as his choices of cultural-area units and his recurring assumption that each of these cultural areas repertoire can be represented by a single song or style [170].

In a much more recent study, Savage et al. present the results of an analysis in which the aspirations of Lomax are strongly echoed. A carefully curated sample of music recordings is examined for musical universals, properties of music that can be found in each of the recordings [171]. Using comparative methods from evolutionary biology, historical relationships between related cultures are controlled for. Though no absolute universals were found among the 32 features that were tested, many *statistical universals* were found, indicating that there are indeed properties of music that apply to 'almost' the entire sample of vocal and non-vocal music. These include Lomax' and Grauer's definition of a song as a 'vocalization using discrete pitches or regular rhythmic patterns or both'. In addition to the statistical universals, eight 'universal relationships' between musical features are identified, i.e., pairs of features that consistently occur together. All these pairs are connected in a network that centers on *synchronized group performance* and *dancing*. The network also contains features related to *drumming* but, somewhat surprisingly, excludes pitch-related features. Altogether, the results are read as the first confirmation of a recent hypothesis by Fitch [52], proposing song, drumming, dance and social synchronization as the 'four core components of human musicality', and a starting point for future cross-cultural comparisons of musical features.

In popular music, Schellenberg's study on emotional cues in a sample of Top 40 records uses manual annotations of two well-established cues of emotion in music, the mode (major vs. minor) and the tempo (fast vs. slow) [173]. These annotations are used to test the hypothesis

that popular music has become more sad-sounding and emotionally ambiguous over time. The dataset of 1010 songs was sampled from the Billboard Hot 100 list, taking the top 40 for each year of the second half of each decade between 1960 and 2010. The tempo was measured by expert assistants who were asked to tap along, from which the tempo was calculated using a software tool. The mode was also determined by experts and defined as the mode of the tonic triad. (Some songs, mostly of the hip hop genre, were considered to have an 'indeterminate mode'.) In the subsequent analysis, it was found that songs have evolved to make use of the minor mode more often over time, with the discrete variable mode accounting for 7.0% of the variance in recording year, and the proportion of minor songs doubling over five decades of data. With regard to tempo, it was found that both major and minor songs have decreased in tempo significantly: major-mode songs by 6.3 beats per minute (BPM) per decade, on average, and minor-mode songs by 3.7 BPM, per decade. The conclusion frames this as an increase in sad-sounding and emotionally ambiguous songs –where emotional ambiguity of a song is equated to minor-mode songs being fast and major-mode songs being slow– because of the stronger change in tempo on major-mode songs. Though the authors cite other research in support of this finding, including a study on the negativity of lyrics, the conclusions in [173] generally point to interpretations that depend very much on one's reading of the prior literature on mode and tempo as emotional cues (most dealing with instrumental music), and how well the conclusions therein carry over to music with lyrics.

3.2.2 *Corpus Analysis Based on Symbolic Data*

Alan Lomax in the 1970's was an early adopter of computers for the statistical analysis of his data. However, with the widespread availability of personal computing at the end of the twentieth century, increasingly complex statistics could be computed. The hand-coding of audio features was no longer necessary, and more advanced computational approaches could now be pursued. This was first and fore-

most an opportunity for those working with the first digitized scores. Hence, symbolic corpus analysis goes back much longer than its audio counterpart, which was made possibly only after audio researchers in music information retrieval disseminated developments first made in speech processing, to the music domain.

In the field of music cognition, like in Lomax's field, several authors have analyzed collections of Western and non-Western music, in search of pervasive, cross-cultural trends. Huron reviews much of his research on this topic in his book *Sweet Anticipation: Music and the Psychology of Expectation*, and in an earlier lecture series on the same topic [79, 80]. Huron reviews theories of expectation by Leonard Meyer, Eugene Narmour (Implication-Realization, I-R), and Lerhdahl and Jackendoff (the Generative Theory of Tonal Harmony, or GTTM) and contrasts them with empirical findings by Henry Watt in the 1920's, Vos and Troost in the 1980's, his student Paul von Hippel and himself, many of them based on corpus studies. Synthesizing these results, Huron proposes a set of five 'robust melodic tendencies', statistical properties of melodies that are shown to hold in various musical cultures. The five patterns include step declination (the tendency of large intervals to go up and small intervals to go down), melodic regression (the tendency of melodies to return to the median pitch) and the melodic arch.

Conklin and Witten, in the 1990's, devised a model of music expectation that is entirely based on statistical learning [204]. This *multiple viewpoints* model proposes an account of how multiple representations of a stimulus—series of note lengths, series of pitch classes, series of melodic intervals...—maintained in parallel, each contribute to an estimate of the most probable next event. Ten years later, the multiple viewpoints model was further formalized in terms of information theory by Pearce and Wiggins as the *IDyOM* model of musical expectation. The model is validated in several corpus studies, examining how well specific trends in the corpus can be explained [151].

A frequent collaborator of Pearce and Wiggins, Müllensiefen used a set of symbolic music features to analyze several interesting symbolic corpora, including melodies off the Beatles album *Revolver* [95], and a

set of stimuli used in a music memory experiment [134]. The feature set draws on inspiration from natural language processing (N-gram models and latent semantic analysis in particular), and descriptors first proposed by Huron, among others.

In musicology, harmony has been a popular subject of corpus studies. Two notable analyses were done De Clercq & Temperley [41], and Burgoyne et al. [27]. In [41], De Clercq & Temperley transcribe the chords for a corpus of 99 rock songs, about 20 for every decade between 1950 and 2000, and analyze the transcriptions in terms of chord root transitions and co-occurrence as they evolved over time. In their findings, they highlight the strong (but decreasing) prominence of the IV chord and the IV-I progression. In [27], Burgoyne presents an analysis of 1379 songs out of the *Billboard* dataset of popular songs (the complete set), in terms of chord composition. The compositional analysis centers around the representation of the dataset as a hierarchical clustering of the 12 possible roots, a *balance tree*, derived from each chord roots' occurrence in each of the 1379 songs. The resulting structure is reportedly consistent with De Clerq and Temperley's analysis. The balances, i.e., the log odds ratios of the branches at each node of the tree, and their inter-correlations, are compared to decade of release and popularity. The findings include a trend towards minor tonalities, a decrease in the use of dominant chords, and a positive effect of 'non-core' roots (roots other than I, V, and IV) on popularity.

Other examples of analyses of distributions of symbolic data include the studies of pitch class and scale degree usage by Krumhansl [100]. Examples of corpus-based analyses of rhythmic patterns include Mauch's analysis of 4.8M individual bars of drum patterns sampled from around 48,000 songs, and Volk and Koops' analyses of syncopation patterns in a corpus of around 11,000 ragtime MIDI files [94, 121, 195].

Seeking to connect a musicological interest in Western classical style to perceptual theory, Rodriguez-Zivic et al. performed a statistical analysis of melodic pitch in the *Peachnote* corpus[1] in [166]. The dataset contains music from 'over 65,000 scores', automatically digitized us-

1 http://www.peachnote.com/info.html

ing OMR. A dictionary based on pairs of melodic intervals is used to represent each 5-year period between 1730 and 1930 as a single, compact distribution. $k = 5$ factors are then identified using k-means clustering, four of which are observed to coincide with the historic periods of baroque, classical, romantic and post-romantic music, and can be read as a description of their stylistic properties. The four periods are roughly characterized by, respectively, use of the diatonic scale, repeated notes, wide harmonic intervals, and chromatic tonality.

A number of the above contributions have resulted in toolboxes dedicated to the analysis of symbolic music corpora. Huron worked with the *Humdrum* toolkit and its associated *Kern* representation of scores, both of which are still used and supported.[2] Pearce made a Lisp implementation of the Idyom model available on the Soundsoftware repository[3] and Müllensiefen's FANTASTIC toolbox, written in R, is also available on line.[4] The *Peachnote* corpus can be accessed through an API at `www.peachnote.com`.

3.2.3 *Corpus Analysis Based on Audio Data*

Much of the existing work involving audio corpus analysis has focused on popular music and non-Western music—two big clusters of music for which scores or other symbolic representations are not often a musical work's most authoritative form. (In both of these groups of styles, music notation is typically either unavailable, or only available because a recording has been transcribed.)

In an example of non-Western music analysis, Moelants et al. describe in [130] a procedure of the automatic analysis of automatically extracted pitch histograms. The procedure is applied to a collection of historic African music recordings, and show evidence for Western influence in the use of African tone scales. Also using tone scale analysis, Panteli and Purwins compare theory and practice of scale intonation

2 `http://www.musiccog.ohio-state.edu/Humdrum/`, `http://github.com/humdrum-tools/humdrum-tools`

3 `http://code.soundsoftware.ac.uk/projects/idyom-project`

4 `http://www.doc.gold.ac.uk/isms/m4s/`

in contemporary (liturgical) Byzantine chant. Analyzing 94 recordings of performances by 4 singers in terms of the tuning and prominence of scale degrees in 8 different modes, they find that smaller scale degree steps tend to be increased, while large gaps are diminished [146].

In an example of popular music analysis, Deruty and Tardieu test a number of hypotheses about the evolution of dynamics in popular music [42]. The hypotheses are formulations of a recurring intuition among producers and consumers, hypothesizing a 'loudness war', a speculative trend in which the loudness of pop songs has gradually increased, in a race between producers of new releases to stand out on the radio. In the study, 2400 recordings released between 1967 and 2011, sampled from a list of critically-acclaimed popular music albums, are analyzed in terms of their energy (root mean square energy or RMS), loudness, loudness range (measuring macro-dynamics) and peak-to-RMS factor (measuring micro-dynamics). They conclude that the energy and loudness have indeed increased, and that micro-dynamics have indeed decreased. Macro-dynamics, however, were not found to evolve significantly.

In the domain of music cognition (specifically, embodied music cognition) one recent study uses corpus analysis to identify the acoustic properties of music that affect walking speed (in m/s) [104]. Leman et al. had 18 participants walk freely to a precompiled playlist of 52 songs, all with a fixed tempo of 130 BPM, and measured their walking speed using wireless accelerometers. The acoustic correlates of walking speed were assessed in a two-stage statistical analysis involving feature selection from a candidate set of 190 audio descriptors, and a model selection stage, using the 10 best features, in which the best fitting linear model was found via (nested) cross-validation. The best performing model involves 4 features and is found to explain 60% of the variance after a second ('outer') cross validation. The features are said to capture variations in pitch and loudness patterns at periods of three, four and six beats.

A few contributions in the domain of performance studies have also involved the analysis of a dedicated audio corpus. In [97], for example, Kosta et al. compare loudness dynamics across 239 piano

performances of a selection of 5 Chopin mazurkas. They find that pairs of dynamic markings in the score don't always correspond to an expected change in decibel levels, and expose further non-trivial dependencies between loudness, note density and dynamics.

Finally, two relatively recent studies have focused on the topic of popular music evolution. In [175], pitch, timbre and loudness features are analyzed for a sample of songs, with dates, from the Million Song Dataset (MSD). In [122], songs sampled from the Billboard charts of US popular music are analyzed using techniques from text mining and bio-informatics. Given their topic—popular music—and their relevance to other results presented as part of this thesis, we will review these two studies as a case study in section 3.4.

A note on Automatic Transcriptions

While the music information retrieval community has made substantial progress in its efforts to improve the transcription of audio to symbolic data, considerable hurdles remain [179]. To our knowledge, no corpus analysis studies have yet been proposed that rely on the complete polyphonic transcription of an audio corpus. And understandably so, since the assumptions of the transcription model would have a considerable impact on the quality of the data, and, worse, most certainly introduce biases in the data itself.

One approach that illustrates the inherent risks in the analysis of transcribed corpora, is Barthet et al.'s study on chord data mining [8]. In an analysis of one million automatically transcribed chord sequences, Barthet et al. acknowledge the drawback that is the chord recognition system's error rate. However, they 'assume that the most frequent patterns emerging from the analysis should be robust to noise'. Even if this is the case, no mention of potential structural biases is made. Many of the state-of-the-art chord transcription systems rely on a form of priors that govern in which order the system expects to see chords. For such systems, a simple count of root transitions would already return biased results.

Barthet's chord extraction was performed using Chordino, which only involves frame-by-frame matching of chroma features to a dictionary of chord profiles, followed by 'heuristic chord change smoothing' [120].[5] The system lacks a language model, so it puts fewer restrictions on its output. Nevertheless, it has been trained on or optimized for a particular collection of music, and the patterns present and not present in that particular dataset (e.g., the very popular Beatles dataset) will be reflected in its output. A similar argument applies to other kinds of transcription (e.g. melodic transcription), as well as for Rodriguez-Zivic's study described above, as it relies on OMR for the transcription from images of scanned scores.

Most existing work on audio corpus analysis has therefore focused on the analysis of audio features rather than automatically transcribed melodies or chords.

3.3 METHODOLOGICAL REFLECTIONS

Studying music through the analysis of a collection comes with its own particular set of challenges. What are important issues in audio corpus analysis that are not typical issues in MIR, and how can they be addressed? This section seeks to answer that question by looking at the methods reviewed above, and by reviewing existing commentaries on the use of music information retrieval technologies in interdisciplinary scientific research. Note that several of the points below apply to corpus analysis in general as much as they apply to audio.

To structure the discussion, we start from the observation that a majority of studies follow variations of the same procedure, involving the choice of a research question or hypothesis, a dataset, a feature set and an analysis method. Each of these steps will now be discussed.

5 http://isophonics.net/nnls-chroma

3.3.1 *Research Questions and Hypotheses*

Outside of MIR, the prevailing scientific practice of addressing a research question using data involves hypothesis testing. Generally, a prior intuition or theory is followed so as to arrive at a prediction or *hypothesis*, describing a certain trend. A good theory leads to a hypothesis that can be falsified or *rejected* in a *statistical test*. A statistical test looks at data and decides whether the data contradicts the hypothesis, and the hypothesis must be rejected, or not. If the hypothesis is not rejected, it is not considered proven, rather, there is no evidence that it is wrong. Tests are performed at a certain significance level α, specifying the probability one allows for a trend being found due to chance, i.e., due to a coincidence in the sample. Lowering α decreases the chance of false discoveries (type I error), but increases the chance of rejecting an existing effect (type II error).

Much of this carries over to music research, and has been applied in countless studies, but it bears repeating how vastly different the procedure is from prediction-evaluation paradigm seen in classic MIR, where hypotheses are rarely stated explicitly. It is also important to look at some challenges that come with hypothesis-based research that are not often acknowledged.

In [81], Huron points to the importance of taking care when choosing hypotheses in music research. He stresses that, historically, hypotheses were typically formed before any data could be acquired. With the arrival of large datasets, however, it is tempting to formulate hypotheses based on an initial exploration of ones dataset, causing the data to be used twice. This increases the chance of confirming, in subsequent tests, a trend that was spurious to begin with, an artifact due to sampling that wouldn't be present if new data were collected. Huron therefore advices against such exploratory activities, calling for the treatment of datasets as finite resources that lose value every time a correlation is computed. If needed, exploratory studies should be done with idiosyncratic rather than representative data.

Several of the above corpus studies do not follow a hypothesis-driven approach. They try to answer questions like: 'What are the

salient patterns in this particular genre?' (Mauch, Koops) and 'How does the use of patterns evolve over time?' (Burgoyne, Volk). These are common musicological questions that cannot be formulated in terms of a single hypothesis. Some studies therefore explicitly center on what could be considered exploratory analysis.

Burgoyne, in [21], presents the results of an experiment in which a Bayesian network or probabilistic graphical model (PGM, see also section 4.4) is learned from a set of variables relating to the harmony and chart position of the songs in the Billboard dataset. This approach aims to expose pairwise correlations between variables that are significant after the effect of other variables is removed, without a prior hypothesis as to which of them are expected or why. Leman et al., in [104], use cross-validation to make the most of the data they have collected, as a time consuming experiment like theirs cannot easily be repeated to collect more data.

Because these studies are not strictly followed by a confirmation on new, independently collected data, their approach is at odds with Huron's advice. Are they therefore invalid? Not necessarily, many of these are respected, peer-reviewed results. Exploratory analysis, with proper use of statistics, can be useful and valid if it is accounted for using appropriate significance levels.

This suggests a spectrum of methods practiced, on which Huron's position represents a rather conservative perspective, which allows, when taken to its extreme, only for yes-or-no research questions, and not for questions of the what/when/where kind (e.g., what makes a song popular, or, when did ragtime syncopation patterns change most), or at least not with a single dataset. Analysis methods will be discussed later in this section, but we can conclude for now that prior hypotheses are a valuable, but not the only option, and that several alternative analysis methods have been developed enough for more open-ended questions to be asked and answered, with appropriate precaution, in a statistically rigorous way.

3.3.2 *Choice of Data in Corpus Analysis*

The intricacies of proper dataset curation make for a PhD topic of their own, see e.g., Burgoyne's account in [26] and Smith in [182]. Though referring to a study as 'corpus analysis' may make it seem as if the corpus is a given, that should be analyzed to answer a particular question. Ideally, of course, the choice of dataset will *follow* the research question: the set of musical works is chosen that allows the question to be addressed most reliably. Compiling a dataset generally requires careful demarcation of the kind of music the research question pertains to, and careful sampling to adequately represent this population.

In the context of corpus analysis, it should be stressed that many of the datasets that are used in music computing have *not* been compiled to be representative of a particular music, but to serve as a test bed for various MIR technologies. The content varies accordingly. The Beatles dataset, often used for chord extraction evaluation, contains a wealth of rare and challenging harmonies, but draws on the work of just a single group of artists. The later Billboard dataset is a much more representative sample of popular music, as it is sampled from the Billboard Hot 100 chart. For example, it includes duplicates to reflect the varying number of weeks songs stayed in the charts [26]. But it was also constructed with large-scale harmonic analysis in mind. Furthermore, it contains only music released up to 1991, when Billboard's own measurement strategies changed. As a result, new genres such as Hip-Hop, that cannot be characterized in terms of chords and modes as easily as earlier genres, are missing from the corpus. The result is a potential bias towards songs with harmonies that can be parsed in terms of traditional music theory. Finally, the Million Song Dataset (MSD) was compiled using a variety of criteria: by downloading the music of the 100 artists for each of The Echo Nest's 200 most-used tags, plus any artists reached by a random walk starting from the most familiar ones, according to The Echo Nest.[6] It is explicitly biased towards challenging rather than representative musical material (e.g., by including music relating to an intentionally broad range of

6 see `http://labrosa.ee.columbia.edu/millionsong/faq`

tags) [14]. It follows that current MIR datasets aren't necessarily suitable for corpus analysis as defined here.

In [81], Huron also notes how large datasets, theoretically, allow for trends to be found with low error rates, both of type I and type II. However, this also makes statistical significance of spurious trends due to a biased sample more likely. Therefore, it is always good practice to validate findings derived from a corpus with new, independently gathered data. However, of all suggested practices, this is one of the most difficult and potentially expensive ones. And, as almost all music data are per definition historic, there is often a fundamental upper limit to how much new data can be acquired.

3.3.3 *Reflections on Audio Features*

Just like the datasets used in music information retrieval aren't necessarily appropriate for corpus analysis, audio features can be inappropriate too. One prominent voice of reflection in MIR, Sturm has argued that a lot of studies in the audio-based music information retrieval field have focused excessively on flawed evaluation metrics, resulting in vast over-estimations of the modeling power of many widely used technologies, including audio features.

In [184], Sturm inspects a large number of studies that have all used the GTZAN genre classification dataset, a dataset for genre recognition training and evaluation compiled by Tzanetakis in 2002 [188], as well as the dataset itself, and observed that there is a hard ceiling to the performance numbers that can be realistically obtained. The ceiling is due to mistaken tags, repeated entries, and other issues. Despite this ceiling, several systems report near-100% accuracies. The study then shows how some of these impossible performance numbers can be attributed to errors in the evaluation, while others cannot be replicated at all.

However, rather than putting the blame with the authors for the quality of their contributions, Sturm examines the evaluation pipeline itself, to conclude that the evaluation of classification systems just based on their accuracies, is flawed. In essence, much of the appar-

ent progress as reported using the above dataset, should be seen as fitting systems to the dataset rather than the task, even when cross-validation is used to avoid overfitting.

What does this imply regarding the use of audio features? As a result of the above practices, Sturm suggests, the current state-of-the-art systems in genre recognition do not listen to the music as much as they listen to a set of largely irrelevant factors that turn out to be proxies for the genre labels as they have been assigned in the GTZAN dataset. In short, features that have been shown to do well in predictive, classification-based MIR, aren't necessarily meaningful descriptions of the music. Or, again in other words, it is not because a feature works in a MIR system, that it is meaningful.

Sturm's concern is echoed in some of the arguments made by Aucouturier and Bigand in [6]. Aucouturier and Bigand, an MIR researcher and a cognitive scientist, examined some of the possible reasons for the MIR communities' limited success in gaining interest from music cognition and neuroscience. As one of seven problems they have identified, the authors stress that many of the audio features used by MIR may seem, at first, to have some cognitive of perceptual basis. Yet often enough, they do not. Whereas the use of the spectral centroid as a timbre descriptor can be justified using evidence from psycho-acoustics, spectral skewness, for example, is mostly a convenient extension of the spectral centroid (see Section 2.1.2), rather than a realistic perceptual attribute of timbre. Likewise, MFCC features may build on some perception-inspired manipulations of an acoustic signal, like the use of the Mel scale and the use of a logarithm for compression. But the discrete cosine transform, that is used next in the computational pipeline, is simply unlikely to have a neural analog.

Similarly, Haas and others have noted that a worrisome number of data-oriented MIR systems completely neglect time [29, 61]. They show that the so-called bag-of-frames approach to music description (audio, especially) is very widespread. In this approach, features are computed over short frames, and frames are pooled by taking the mean or variance, or some other summary statistic that is invariant to order. The efforts that have been made to re-introduce time in music

description have largely been focused on symbolic data (see examples in [61]), leaving audio features behind. In general, Haas points to a variety of opportunities in incorporating more musical knowledge and music cognition in music description.

Naturally, this is not to say that everyone has been doing everything wrong. It is rarely the intention of MIR researchers to develop realistic models of neural mechanisms. In a typical goal-oriented setting, features will be used if they help improve the precision of an algorithm, regardless of whether they have a verified psychological or neural underpinning. Aucouturier and Bigand are right to point out that this procedure is fundamentally at odds with scientific practice in the natural sciences, where variables aren't added to a model because they improve its performance, but because they correspond to a theory or hypothesis that is being tested. But, as most MIR researcher would attest, 'science' may just often not be the goal [6].

And then there is another appropriate nuance that isn't often discussed: music cognition and neuroscience are themselves at times divided, on topics such the learned and culturally mediated nature of mental representations [71], and the neural basis of apparent cognitive 'modularities' [158]. Yet, the above illustrates why it is important to exert caution whenever a feature that was originally developed for some MIR application is used in the context of scientific music research, even if it is widely used.

To conclude, existing commentators point to a tendency among researchers to choose convenience and prevalence over relevancy and cognitive or perceptual validity of features. While efforts in feature design have resulted in an impressive canon of powerful audio features, most are *a priori* uninformative, and therefore of little use in interdisciplinary research. There is a lot of room for the perceptual validation of existing features and the design of novel cognition-inspired features that better align with cues that are known to be important in human music perception and cognition.

3.3.4 *Reflections on Analysis Methods*

Similar arguments can be heard in discussions on the analysis and learning algorithms that integrate features to make predictions. Aucouturier and Bigand, in the second out of their seven problems, criticize the algorithms used in MIR with a very similar argument as they first made about features: algorithms are presented as if they reflect cognitive mechanisms, but they do not. Even if all the features in a model were accurate measurements of plausible perceptual or cognitive correlates of a musical stimulus, most of the commonly used statistical models wouldn't reveal much about how these attributes are combined into more complex judgements on e.g., the perceived valence of the emotions conveyed by the music (as in the example given in the article). Even if a feature is only incorporated into an algorithm if its individual predicting power is tested and validated, it may be unclear "what sub-part of a problem that feature is really addressing", especially when modeling a highly cognitive construct like genre or emotion [6].

To Huron, statistical practices are a recurring concern. Along with his recommendations on choice of hypotheses and data (reviewed above), he reviews the statistical caveats that come with the use of big datasets in musicology [81]. As it has become easier than ever before to undertake a large number of experiments, thresholds of significance should adapt: if, in a typical study, well over 20 relationships have been tested for significance, some will end up being spurious, and a significance level well below the traditional 0.05 should be considered. The Bonferroni correction for multiple tests, an adjustment of the significance level α based on the number of tests, is traditionally used to address this issue:

$$\alpha_B = 1 - (1 - \alpha)^{1/n} \approx \alpha / n \tag{21}$$

with n the number of tests and α the overall significance level of the study (e.g., 0.05). It is an effective measure against overfitting to a sample. Usually, however, n only counts formal tests. When exploratory processing of a dataset is involved, α_B should reflect the substantial

amount of visual exploration and eyeballing of potentially interesting relationships that is often done prior to any formal testing. And as others have argued in the debate on the use and misuse of *p*-values in science, reporting effect sizes also helps to communicate results convincingly: as datasets get bigger, it is increasingly easy to find significant, but small effects [81].

In standard machine learning-style analysis, significance is even more difficult to asses. An SVM classifier may tell you if a feature is helpful or not, but it doesn't reliably quantify the significance of that feature's contribution, let alone its effect size. It is often even unclear in what 'direction' a feature contributes—suppose a classifier predicts a strong influence of tempo on whether or not a song is perceived as happy, it often won't tell you what range of tempos make a song more happy, especially if the classifier is a 'distributed' model like those based on boosting, forests or neural networks.

Another statistical modeling issue that is not typically deemed relevant in machine learning is the role of variable intercorrelations and confounding effects. A confounding effect occurs when some trend is attributed to one variable while it is in reality due to another (observed or unobserved) variable. When dealing with a set of correlated features, it may turn out that some of the features that correlate with a dependent variable of interest, contribute little explanation in the presence of other features; they are, as is said, 'explained away'.

An ideal statistical analysis that is focused on not just correlations, but on 'effects', allows to control for obvious confounding correlations, and acknowledges the possible effect of correlations that couldn't be controlled for. This kind of 'causal modeling' is far from trivial. The common perspective is that it requires *interventions* that allow some variable to be willfully adjusted, as in a randomized controlled trial. When dealing with historic data, such as any music collection, intervention is not typically an option, and the feasibility of causal modeling can be disputed.[7] Others have argued that, under certain restrictions, causal relationships may still be obtained. Probabilistic graph-

7 See, e.g., Sturm's thoughts on the subject `http://highnoongmt.wordpress.com/2015/07/30/home-location-and-causal-modeling/`

ical models, for example, model conditional relationships between variables, allowing for a certain amount of insight into the confounding effects between the observed variables [93]. Most other statistical methods, however, don't model the effects of individual variables, and therefore don't account for interactions in the causal sense.

Studies with many variables, hypothesis- or discovery driven strategies alike, face an additional statistical challenge: the amount of data required to fit a reliable model may scale unfavorably with the number of variables. This problem is sometimes referred to as the 'curse of dimensionality'. It denotes a wide range of issues that arise because data get sparse quickly as the dimensionality of the feature space goes up. Mathematically: distances between uniformly distributed data points in a high number of independent dimensions tend to lie mostly on a relatively thin shell around any given reference data point, and data points tend to have a very similar degree of dissimilarity to each other [66].[8]

This forms a recurring challenge in statistics and statistical learning. Many statistical models involve $O(m^2)$ parameters. An arbitrarily structured multivariate normal distribution in m dimensions, for example, requires $m^2 + m$ parameters to be fit. Any fit with less than m m-dimensional data points will therefore fail, as there are more degrees of freedom in the parameters than in the data points. A good fit will take several times that number. This dependency will show up in any model that acknowledges correlations, i.e., any model that doesn't treat its dimensions as completely independent—a very strong assumption, most of the time.

The problem gets worse for models that account not just for binary interactions between features, but between any combination of variables, e.g. learned graphical models. Because the number of dif-

8 The technical argument is given by Hastie as follows: for inputs uniformly distributed in a m-dimensional unit hypercube, "suppose we send out a hyper-cubical neighborhood about a target point to capture a fraction r of the other observations. Since this corresponds to a fraction r of the unit volume, the expected edge length will be $e(r) = r^{\frac{1}{m}}$. In ten dimensions: $e(0.01) = 0.63$ and $e(0.10) = 0.80$, when the entire range for each input is only 1. So, to capture 1% or 10% of the data to form a local average, we must cover 63% or 80% of the range of each input variable."

ferent graph structures grows super-exponentially with the number of nodes, many tests are typically required to find the best candidate, even if heuristics are used [16]. All of these issues, while less cumbersome in machine learning and prediction, put unfortunate limitations on the complexity of statistical analyses of high-dimensional data.

While this last point reads like an argument against large-feature-set analyses at large, there are exceptions and alternatives. For example, if the number of data points is low or the number of dimensions very high, heuristics and regularization may be used. In the specialized literature, techniques for the estimation of structures in sparse, correlated datasets are emerging, e.g., regularized covariance matrix estimation [15]. If the number of data points is sufficient for a reliable estimate of the covariance matrix, dimensionality reduction (e.g., PCA) may be applied to facilitate further model selection, ideally accompanied by steps taken to avoid compromising interpretability in the process.

Overall, choosing a good modeling approach seems to elicit the same problems as choosing the right audio features: models that have been shown to work fall short on requirements that are inessential for information retrieval but crucial in scientific modeling, including simplicity (in terms of the number of parameters), accounting for correlating variables, and accounting for multiple tests. Furthermore, the analogy of analysis methods with perceptual and cognitive modularities is often flawed. When, finally, a statistical model seems appropriate, there may not be enough data for the model to be fit, or for any true effects to surface after significance levels are adjusted to reflect all testing involved in fitting the model.

3.4 CASE STUDY: THE EVOLUTION OF POPULAR MUSIC

Having reviewed a range of issues that come up in the corpus analysis of audio data, we present a case study that illustrates how some of these issues have been addressed in practice, and others have not.

The case study focuses on two publications on the topic of popular music evolution, by Serrà et al. in 2012, and Mauch et al. in

2015 [122, 175]. The common question they address is, roughly, 'has Western popular music, over the past decades, become more or less diverse'? The studies are of particular interest because of their focus on popular music, but also because they arrive at partially contradicting conclusions on the supposed decline of diversity in popular music. Both studies are also the work of researchers with a considerable authority on the subject of audio analysis, and have been given wide attention in the popular press.[9]

First, each study's methods and results will be summarized and compared. A discussion section will then discuss the differences between the studies in terms of data, features and models. Along the way, some additional pitfalls that haven't been brought up in the methodology literature will be identified. Finally, the most important differences and pitfalls will be summarized at the end. A side-by-side comparison of the studies' analysis pipelines is given by the diagram in figure 16.

3.4.1 *Serrà, 2012*

In the first study, by Serrá et al., pitch, timbre and loudness features are analyzed, to answer a number of questions that includes the one above [175]. The dataset is a sample of 464,411 songs from the MSD, all released between 1955 and 2010. The features correspond to precomputed pitch, timbre and loudness features as provided by The Echo Nest[10], computed over 10 million consecutive frames for every year of data, sampling from a five-year window. For each feature, a codeword dictionary is then extracted, yielding a vocabulary of pitch, timbre and loudness codewords for each year.[11] The studies hypothe-

9 E.g., `http://graphics.latimes.com/music-evolution-hip-hop-rap/`, `http://www.theguardian.com/music/2012/jul/27/pop-music-sounds-same-survey-reveals`

10 `the.echonest.com`

11 Codewords are simplified, discrete representations of multidimensional feature vectors. The mapping of feature vectors to codewords is often found by applying clustering to a dataset, after which each data point is mapped to the closest cluster center. Here, a simpler heuristic was used to discretize the features [175].

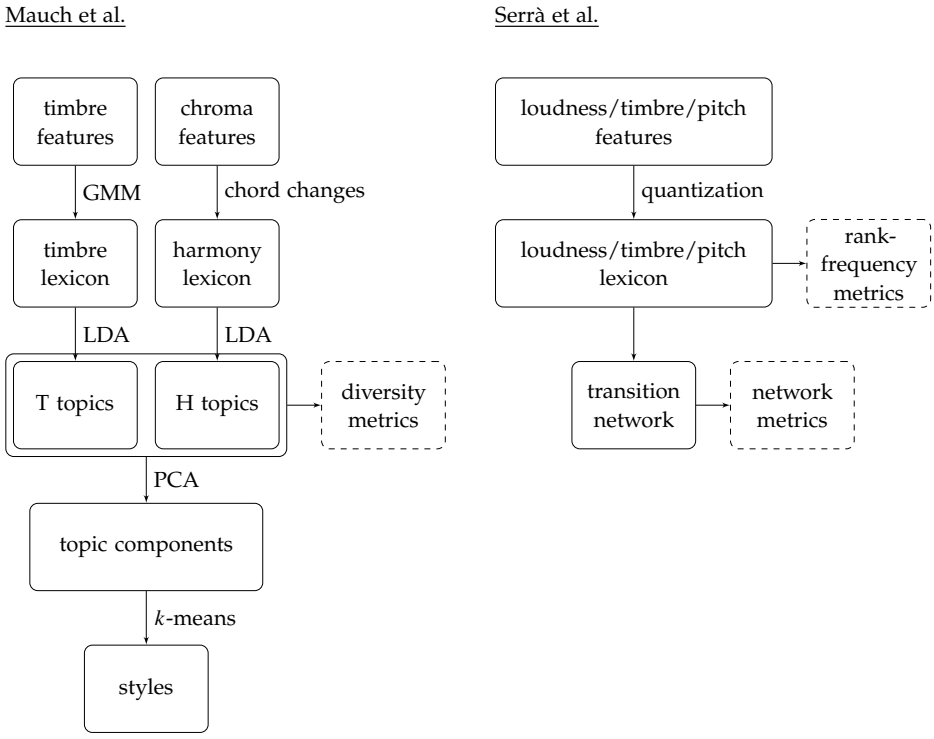


Figure 16.: Diagram comparing Mauch's and Serrà's analysis approaches.

sis questions are addressed through a statistical analysis of the distribution and transition network of these codewords.

In a short analysis prior to computing loudness codewords, the empirical mean of loudness values is found to have increased, from -22dB_{FS} to -13dB_{FS}, or about 0.13dB_{FS} per year, while dynamic range hasn't evolved significantly—findings that are loosely consistent with Deruty et al.'s later results described in section 3.2. In terms of loudness codewords and transitions, the network's topology is maintained throughout the decades.

When timbre is analyzed, the codewords are modeled using a rank-frequency distribution based on Zipf's law. Zipf's law states that an event's frequency of occurrence can be modeled as a logarithmic function of its rank when all events are sorted by frequency. The rank-frequency distribution that is used is parametrized by the exponent parameter β. For timbre codewords, it is found that β decreases over time since 1965, indicating a homogenization of the timbre palette. Like loudness codewords, no changes in the topology of timbre codeword transitions could be found.

In the pitch domain, no change in β is found. However, some trends have been discerned when looking at the pitch transition network. The most obvious indicator of diversity of transitions, the median degree of the network k, is unchanged. However, the clustering factor C, assortativity Γ and the network's average shortest path l are found to change significantly, in a way that, together, constitutes a decrease in 'small-worldness' of the network, showing a restriction of the possible transitions, and thus a decrease in the variety of observed transitions [175].

To sum up, Serrà et al. find a progressive increase in the predictability of pitch use, a tendency towards mainstream sonorities in the timbre domain, and an increase in loudness of productions.

3.4.2 *Mauch, 2015*

Mauch et al. follow a very different approach, leading to a rather different conclusion, in another quantative analysis of popular music evolution [122].

Like Serrà, Mauch et al. analyze audio data for a large corpus of popular music sampled from the last 50 years. Specifically, their corpus extends from 1960 to 2010 in 17,094 30-*s* segments from the same number of songs. Instead of Echo Nest features, freely available tools were used to compute chroma and timbre descriptors from these segments. The study also employs a quantisation step to convert each of these segments to a sequence of *words* out of a newly constructed lexicon, or in this case, two: one for timbre and one for harmony. The timbre lexicon is obtained using unsupervised clustering based on Gaussian mixture models (GMM), of the 14-dimensional audio feature space (12 MFCC coefficients, one delta-MFCC coefficient, and the zero-crossing-rate). In GMM, data points (here: frames) are assigned to an optimized number of Gaussian-shaped clusters. The optimal number number of clusters is found at 35. The harmonic lexicon consists of 192 possible intervals between the most common chord types.[12].

However, the text data approach is taken further than it is in [175]. With each 30-s frame of audio converted to a single word, *topics* are extracted from the data, using a topic modeling technique from text mining called latent Dirichlet allocation (LDA). This hierarchical model regards documents as distributions over a set of topics, which are themselves distributions over the lexicon. 16 topics are found, 8 for timbre and 8 for harmony. Following an analogy with evolutionary biology, these topics can be regarded as *traits* or *character expressions* associated with the 'genome' that is, in this study, the string of words of which a song is composed.

The documents' distributions over these topics are used to compute four measures of genetic diversity, borrowed from bio-informatics. The diversity measures show substantial fluctuations over time, most

12 Four modes (major, minor, major 7, minor 7) × four modes (for the second chord) × twelve root intervals = 192, plus one label for ambiguous harmonies [122]

notably a drop in the early 1980's, followed by an increase to a maximum in the early 2000's. Interestingly, however, no evidence is found for the progressive homogenization of the music in the charts as posed by [175], neither in the timbre domain, nor in the harmony domain.

In a second and third part of the study, one more layer of abstraction is added when the topics distributions are grouped into *styles*, similar to populations in genetics. First, the topic space is further reduced to 14 dimensions, using PCA with standardization of the components. The styles are then found using k-means clustering on the timbre and harmony topics. The best fitting number of styles is found to be 13. At each stage, the feature spaces employed in this paper, be it the lexicon, the topic space, or the musical styles, are checked for interpretability. This is achieved using a combination of expert annotations on a representative mixture of short listening examples, interpretation of the chord labels by the authors, and enrichment of the styles with tags from *Last.fm*.[13]

The second question to be addressed is: when did popular music change the most? To address this, the 14-dimensional topic space is first used to compute a distance matrix over all analyzed years, and a novelty curve can be computed. A novelty function, as introduced in Section 2.2.2, tracks discontinuities in the time series. It is found that three years brought significant change to the topic structure of popular music: 1991, 1964, and 1983, of which the one in 1991 is the biggest. Using a similar analysis of the 13 extracted 'styles', it is found that these years coincide with the moments that soul and rock took over from doo-wap (in 1964), the year that soft-rock, country, soul and R&B made place for new wave, disco and hard rock (in 1983), and the rise of hip hop-related genres to the mainstream in 1991.

3.4.3 *Discussion*

The two studies above have both received wide exposure in the specialized and popular press, despite conflicting conclusions on the supposed decline of diversity in popular music. Where does this contra-

13 `www.last.fm`

diction arise? We discuss the most important differences between the studies in light of the concerns raised in the previous section (section 3.3). We distinguish between choices of research questions, data, audio features and analysis methods, and focus on the common part of each study's research question: is there evidence that popular music has become more homogeneous, more predictable or less varied over time?

Data

The first critical difference is the music sample of choice. Mauch et al. aim to use the complete Billboard Hot 100 as their sample and manage to include about 86% of the complete list. Serrà et al. choose to construct a sample that includes a large portion of the Million Song Dataset.

Both datasets have their advantages and drawbacks. The MSD is much larger, but sampled rather arbitrarily (see earlier). Therefore it is neither controlled for popularity, nor a complete picture. Neither Serrà or Mauch discuss the option of controlling for popularity. A popular music corpus, ideally, gives more weight to songs that were listened to more often. Mauch do this to a primitive extent, by sampling only songs from the Billboard Hot 100, but much more could be done: the sampling procedure used to compile the Billboard dataset by Burgoyne, for example, allowed for songs to be included several times if they stayed in the charts longer [26].

But the Billboard Hot 100 also has other flaws. For instance, it is known that in 1991, the method of measuring popularity as a function of radio play and sales, was automated, and as a result, drastically changed [26]. This calls for some caution when interpreting the claim that popular music's biggest moment of change came in 1991—one should at least consider the possibility that this effect is in fact, sample noise, an effect due to the way the Billboard Hot 100 list was compiled. The paper, making no mention of the measurement procedures of the Billboard organisation, does not address that possibility. It goes to

show that, as discussed above, a consistent sampling strategy is crucial in corpus-based studies.

Thus, neither of the studies seem to have properly considered the issues we formulated regarding datasets for corpus analysis, though Mauch et al. make a somewhat stronger case by not choosing the charts over the Million Song Data. The different approaches and outcomes are reminiscent of several anecdotes that are used to illustrate a common 'big data' fallacy, in which a sample is deemed reliable because it comprises almost all of the population, and its biases are dismissed as if the dataset's size could somehow make up for it.[14]

Features

The studies also differ in their choice of features and statistical measurements on descriptors. While Serrà et al. focus on networks of transitions between code words, Mauch et al. group them into topics and look at the evolution of those topics. While it may seem that Mauch et al. disregard the time component that is very often overlooked, but somewhat included in Serrà's analysis, it must be noted that, in the latter, changes in the transition network's topology only drive the homogenization effect in the pitch domain, not in timbre or loudness. Furthermore, Mauch et al. effectively do include some time information in their representation of pitch, as the harmony features used to extract the harmonic topics are based on chord transitions rather than just the chords themselves.

Other description-related differences remain. The descriptors used by Mauch do not include melody, whereas the features used by Serrà arguably could, and Mauch narrows harmony down to a space defined in terms of chords (and specifically: triads), which, as other have noted, are perhaps not appropriate for the description of recent 'urban' music (hip hop and related genres) [26, 173]. On the other hand, Serrà et al.'s network representation only considers binary counts: whether

14 see, e.g., the often recounted case of Literary Digest's predictions for the 1936 US presidential elections. Their poll, one of the largest in history at the time, failed in the end as the result of a bias in both sample and response [183].

or not a code word or transition appears, regardless of its proportion in the sample.

Analysis Methods

A third set of differences and potential issues, ultimately, appear in the analysis methods. The studies use a different set of diversity measures: network statistics (Serrà) vs. bio-informatics measures (Mauch), of which only Mauch's have been validated in other studies with similar research questions.

Serrà's method also runs into an issue related to confounding variables. In his study, both loudness and timbre are reported to homogenize over time. One obvious question that is not addressed is: does increased loudness not affect the range of possibilities left in the timbre palette? If a substantial amount of the timbre-related trend can be explained by an increase in loudness, the results of the paper would look quite different. The conclusions do not acknowledge this. Mauch et al. don't run into this issue, because no trends are found in either the timbre-related set of topics, or the harmony-related set.

Finally, Mauch et al. look for trends only after transforming the code word representation of their data set to the strongly dimensionally-reduced abstraction that are the topics. The study could have included some measurements of diversity on the codeword representation itself, to see if a homogenization can be observed earlier in the analysis.

3.4.4 *Conclusion*

The above analysis brings up a range of substantial differences between the two studies, of which choices in data and analysis methods seem the most salient. Serrà et al. primarily expose their approach to criticism by working with an uncontrolled sample, and by not controlling for loudness in their analysis of the evolution of timbre (and vice versa). Mauch et al. work with a sample that is more convincing, but similarly lacking some control over what makes exactly makes it representative, due to the procedures by which the Billboard charts

are compiled. It is impossible to know which of these differences contribute more to the discrepancy in conclusions without running experiments to test particular variations of their methods. But together, such differences could explain some or all of the disagreement in the results. Conclusions on which approaches should have been followed instead, if any, won't be made here. Section 3.5, however, will list some general recommendations distilled from the observations made above.

On a positive note, both studies deal very thoroughly with a host of other issues: they start from a clear hypothesis, thoroughly motivate their analyses, and refrain from making claims on the cause of the observed effects. In addition, both studies acknowledge potential limitations of their conclusions, e.g. as exemplified by Mauch et al.'s comment that their conclusion is limited to the features they have studied, and that their measures only capture a fraction of the actual complexity of the music in their dataset.

A Note on Interpretability

We close the case study with an observation about the reception of each studies' results, in the general press and among researchers in related fields. Between the two studies, there is large gap in the amount of effort spent on interpretation of the audio features and their abstractions used in the models. Serrà et al. use Echo Nest features, which is a proprietary technology for which the mathematical specifications haven't been published. The abstractions used in the model aren't qualified in terms of musical domain language, but in terms of network statistics. Meanwhile, Mauch et al. use openly available features and collect human annotations for each of the topics in their model, and social tags for each of the styles.

In following the broader reception of both articles and in discussions among colleagues, this discrepancy became especially apparent. Results are seen as uninformative if they are the result of a method that is convoluted or opaque. In contrast, the importance of interpretation of audio features is not widely discussed in the methodology

literature reviewed above. This suggests that interpretable audio features and analysis methods are perhaps more important than authors in the field acknowledge.

Problematically, however, *interpretability* is not easily defined. What constitutes the interpretability of, for example, an audio feature? If we were to define it, we could say it is a feature's property of having an agreed-upon interpretation, where an *interpretation* is an unbiased and sufficiently detailed mapping from the signal or computational domain to natural language or domain language (perceptual, cognitive or music-theoretic). In other words, features that can only be interpreted in terms of computations on the audio signal, carry no information outside of the computational domain, whereas a properly informative feature allows to translate a mathematical trend or pattern into natural language or domain language information.

As a definition, this is rather subtle. MFCC features, as a whole, have an agreed-upon, empirically validated correlation with some subspace of Western musical timbres [186]. Yet, individual MFCC coefficients have no particular interpretation: it doesn't mean much for one or more coefficients to be high or low to most people (except for, perhaps, the first one, a correlate of energy). Moreover, whether or not a feature or analysis method is interpretable is inherently subjective, depending very much on the background knowledge of the audience the results of a study are reported to.

In short, feature interpretability is a quality that is generally considered helpful and important. Yet it cannot be prescribed in exact, universal terms, in part because it is very difficult to define, in part because it is highly dependent on context and audience. It is mostly useful as a predictor for the degree to which a research result may convince scholars outside its immediate domain. We adopt a very pragmatic stance: researchers should adopt the methods that best allow them to communicate with the audience they wish to persuade. Mauch's study, therefore, does have the added benefit, over Serrà's, of offering feature interpretations at every step of the analysis, thus making a stronger case for its results despite the high degree of abstraction of its representation.

3.5 SUMMARY AND DESIDERATA

From the review of corpus analysis research in section 3.2, the methodological reflections outlined in section 3.3, and from the above case study, we distill a number of desiderata. These are desired properties of audio features and analysis techniques, for the context of audio corpus analysis.

3.5.1 *Research Questions and Hypotheses*

Preceding any analysis is the choice of research question. The literature review above roughly leaves room for three options:

1.
   - an external hypothesis, established before any data are collected, or
   - a hypothesis based on an idiosyncratic sub-sample of the data, or
   - no hypothesis, but an explicit strategy for exploratory or discovery-based analysis

with significance levels set accordingly.

3.5.2 *Data*

When choosing a corpus, it is crucial to aim for

2. a representative dataset, carefully sampled from a clearly defined population.

Many existing MIR datasets haven't been compiled to represent anything in particular, and should be used with care, or not at all. When compiling one's own dataset, representative sampling should prioritized over dataset size.

3.5.3 *Audio Features*

To guide the choice of audio features, we put three criteria forward, taken from several of the sections in this chapter.

3. robust features: features can be reliably computed for the entire corpus.

Features that are still very difficult to accurately compute from audio include the precise onsets of a vocal melody or any other source in a complex polyphonic mix, as well as any features based on this information (e.g. inter-onset intervals), and any features that rely on complete transcription of the piece.

When features cannot be treated as independent, it is wise to work with

4. a total number of feature dimensions that stays well below the size of the dataset.

Having a feature set that is easy to oversee aids the transparency of the analysis. However, a modest feature set also helps to avoid the 'curse of dimensionality' as explained in section 3.3.4. Larger feature sets may still be useful for audio corpus analysis, e.g., in conjunction with a dimensionality reduction that allows a meaningful interpretation, and is robust and stable.

Finally, the kind of insight that can be obtained through corpus analysis depends not only on the quality, but also on the perceptual validity or interpretability of features. We should therefore favor:

5. informative features: features have an agreed-upon and validated natural language or domain language interpretation that is accessible to the intended audience of the study.

Ideally, only features are used that are empirically demonstrated correlates of some one-dimensional perceptual, cognitive or musicological quantity. Any summary statistics should be robust and interpretable as well.

3.5.4 *Analysis methods*

Regarding analysis methods, we specify four main desiderata, as guiding principles in selecting an adequate statistical analysis method. First of all, a good model that generalizes to unseen data requires

6. a strategy to avoid overfitting to the sample

by ensuring that no data gets re-used.

Analyses with an interest in estimating causal relationships, require

7. a model that accounts for correlations between measured variables.

See also our earlier comment on Serrà's treatment of timbre and loudness as independent in section 3.4. Consequently, a good analysis also includes as many of the potentially confounding variables as possible, without compromising on its ability to test all resulting interactions.

Analyses with an interest in quantitative relationships, require

8. a model that explains how much each feature contributes.

An ideal quantitative model explains, for each feature, if it contributes positively or negatively, which of the features contribute more, and in the most perfect circumstances: each feature's absolute effect size.

It should be clear by now that such information is the most difficult to obtain. Effect sizes can only be trusted if the underlying model is reliable, which in turn requires all features to be reliable and all potentially confounding interactions to be accounted for.

Finally, when discussing results, it is essential to

9. acknowledge potential issues with all of the above constraints.

e.g., shortcomings of the features, the possibility of not having observed important factors, the possibility of having too many variables or not having seen enough data, and assumptions on the distributions of variables. Any results must be read with care, and effect sizes more so than anything else.

3.6 TO CONCLUDE

Compared to studies with symbolic music data, advances in music description from audio have overwhelmingly focused on ground truth reconstruction and maximizing prediction accuracy, and only a handful of studies have used audio description to learn something about the music.

In this chapter we defined corpus analysis as the analysis of a music collection with the aim of gaining insights into the music itself. We reviewed the most import work in corpus analysis, and the most relevant literature on the subject of modeling music with audio data. Based on this review and a case study of two analyses of popular music evolution, we proposed several guidelines for the corpus analysis of audio data. In short, every step in the choice of hypothesis and dataset and the construction of the feature set and analysis pipeline should be considered carefully. To do this well, a good understanding of the perspective of cognitive science and statistics is desirable.

The above recommendations should be a first step towards this goal. These are not definitive guidelines, but suggestions based on the most relevant literature and an in-depth analysis of two example studies. In the following chapters of this thesis, we will aim to extend the set of available tools that satisfy these criteria.

Part II

CHORUS ANALYSIS & PITCH DESCRIPTION

4

CHORUS ANALYSIS

This chapter presents a corpus analysis of the acoustic properties of the pop song chorus. We address the question: what makes a chorus distinct from other sections in a song?

Choruses have been described as more prominent, more catchy and more memorable than other sections in a song. Yet, in MIR, studies on chorus detection have always been primarily based on identifying the most-repeated section in a song.

Instead of approaching the problem through an application-centered lens, we present a first, rigorous, analysis-oriented approach.

4.1 INTRODUCTION

4.1.1 *Motivation*

The term *chorus* originates as a designation for the parts of a music piece that feature a choir or other form of group performance. In the popular music of the early twentieth century (e.g., Tin Pan Alley and Broadway in New York), solo performance became the norm and the term chorus remained in use to indicate a repeated structural unit of musical form. The same evolution was observed in European entertainment music [129].

In terms of musical content, the chorus has been referred to as the "most prominent", "most catchy" or "most memorable" part of a song [50] and "the site of the more musically distinctive and emotionally affecting material" [129]. It is also the site of the refrain, which

features recurring lyrics, as opposed the more variable 'verse'. While agreement on which section in a song constitutes the chorus generally exists among listeners, attributes such as 'prominent' and 'catchy' are far from understood in music cognition and cognitive musicology [69].

This points to at least two motivations for a deeper study of the particularities of choruses. First, the chorus is a central element of form in popular music. In analyzing it we may gain insight into popular song as a medium, and conscious as well as unconscious choices in songwriting. The concept is also rather specific to popular music, so it may tell us something about where to look for the historical shifts and evolutions that have resulted in the emergence of a new musical style. Second, choruses may be related to a catchy or memorable quality, to the notion of hooks, and perhaps to a more general notion of cognitive salience underlying these aspects. The nature of choruses may indicate some of the musical properties that constitute this salience, prominence or memorability.

Recently, as a frequent subject of study in the domain of music information retrieval, systems have been proposed that identify the chorus in a recording; see also section 2.2.2. Yet, as we will show in the next section, the chorus detection systems that locate choruses most successfully turn out to rely on rather contextual cues such as the amount of repetition and relative energy of the signal, with more sophisticated systems also taking section length and position within the song into account [50, 57]. This suggests a third motivation for the proposed analysis: the potential to advance MIR chorus detection methods with a more informed approach.

The central research question of this analysis is therefore:

> In which measurable properties of popular music are choruses, when compared to other song sections, musically distinct?

4.1.2 *Chorus Detection*

Existing work on chorus detection is strongly tied to audio thumbnailing, music summarization and structural segmentation. Audio

thumbnailing and music summarization refer to the unsupervised extraction of the most representative short excerpt from a piece of musical audio, and often rely on full structure analysis as a first step. The main ideas underlying the most important structure analysis methods are described in section 2.2.2. A more in-depth review of relevant techniques is given by Paulus et al. in [150].

Definitions of the chorus in the MIR literature characterize it as repeated, prominent and catchy. Since the last two notions are never formalized, thumbnailing and chorus detection are essentially reduced to finding the most often-repeated segment or section. A few chorus detection systems make use of additional cues from the song's audio, including RefraiD by Goto and work by Eronen [50, 57]. RefraiD makes use of a scoring function that favors segments C occurring at the end of a longer repeated chunk ABC and segments CC that consistently feature an internal repetition. Eronen's system favors segments that occur $1/4$ of the way through the song and reoccur near $3/4$, as well as segments with higher energy. In most other cases, heuristics are only used to limit the set of candidates from which the most frequent segment is picked, e.g., restricting to the first half of the song or discarding all segments shorter than 4 bars.

Some efforts have also been made in labeling structural sections automatically [148, 149, 205]. Xu and Maddage rely on a heuristic which 'agrees with most of the English songs', imposing the most likely of three typical song structures on the analyzed piece [205]. However, as Paulus and Klapuri show, the datasets that are typically used in structure analysis do not support the claim that a small number of structures recur very often [149]. In the *TUTstructure07* dataset for example, 524 out of 557 pop songs have a unique structure.

Paulus and Klapuri use a Markov model to label segments given a set of tags capturing which segments correspond to the same structural section (e.g., ABCBCD) [148, 149]. This approach performs well on *UPFBeatles*, a dataset of annotated Beatles recordings, and fairly well on a larger collection of songs (*TUTstructure07*).[1] An n-gram

1 Dataset descriptions and links at `http://www.cs.tut.fi/sgn/arg/paulus/structure.html`

method with $n = 3$ and a variable-order Markov model come out as the best techniques. The same methods have also been enhanced by using limited acoustic information: section loudness and section loudness deviation [148]. This boosts the best performance (in terms of per-section accuracy) by up to 4 percent for *TUTstructure07*. Whether the model could be improved with more acoustic information remains an open question.

4.1.3 *Chorus Analysis*

The difference between the present investigation and the chorus detection methods above is both in the goals and in the execution. While chorus detection systems are built to locate the choruses given unsegmented raw audio for a song, this investigation aims to use computational methods to improve our understanding of choruses. And while the computational methods used in chorus analysis relate mostly to structure analysis techniques reviewed in section 2.2.2, we follow a corpus analysis approach—as described in chapter 3. Because structural boundary detection is not part of our goal, we can start from reliable manual annotations of the structural boundaries of a song.

We study the notion of chorus in two corpora: a newly created dataset of early Dutch popular music from the first half of the 20th century, and a large dataset of Western popular music from the second half of the 20th century. The focus on early Dutch choruses was included because it allows us to zoom in on a time period in which popular song developed as a style, and because of the interests of the Meertens Institute and the Institute of Sound and Vision (see section 1.1.2). The more recent dataset allows us to look at trends at a larger scale.

To find trends, we compile a list of appropriate features and model how they correlate with section labels in a collection of song sections. Expert structure annotations for the two datasets allow to parse audio descriptors (see section 4.2.2), into per-section statistics. The analysis of the resulting variables will be presented in sections 4.3 and 4.4.

The contributions of this chapter include the introduction of the 'chorusness' concept, a corpus analysis method to model it, and the resulting model, which we believe can serve MIR applications, popular music understanding and popular music perception and cognition.

4.2 METHODOLOGY

4.2.1 *Datasets*

The Dutch50 *Dataset*

The first dataset, the *Dutch50* dataset, was created especially for this study, and was conceived as a diverse and representative sample of the Netherlands' popular music as it sounded before the 1960s. The *Dutch50* dataset contains 50 songs by 50 different artists, all dated between 1905 and 1951. Figure 17 shows a histogram of the songs' year of release as provided by the publisher. The songs were obtained from compilation releases by the Dutch Theater Institute,[2] acquired by the Meertens Institute. Recurring styles include cabaret, colonial history-related songs, advertisement tunes released on record and early examples of the *levenslied* musical genre [92]. An expert on early Dutch popular music was consulted to validate the representativeness of the selected artists. Structural annotations were made by the author, indicating beginning and end of sections and labeling each with a section type chosen from a list of seven (intro, verse, chorus, bridge, outro, speech and applause).

The Billboard *Dataset*

The *Billboard* dataset is a collection of time-aligned transcriptions of the harmony and structure of over 1000 songs selected randomly from the *Billboard* 'Hot 100' chart in the United States between 1958 and 1991 [26]. The annotations include information about harmony, meter, phrase, and larger musical structure. The *Billboard* dataset is one of

2 http://www.tin.nl

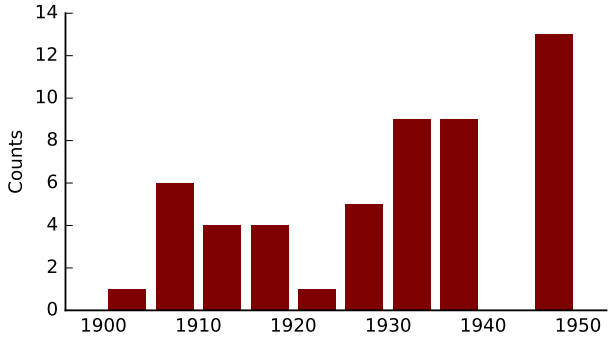


Figure 17.: The distribution of the years of release for the *Dutch50* dataset.

the largest and most diverse popular music datasets for which expert structure annotations exist and one of few to be consistently sampled from actual pop charts. It can be expected to reflect important commonalities and trends in the popular music of the period of focus. It includes a wide variety of popular music subgenres, and suits the goal of drawing musicological conclusions better than other datasets discussed so far, as it is representative of a relevant 'population' (the US charts), and carefully sampled from that population.[3]. This makes it the best available dataset for analysis of popular music choruses. For the present study, the complete v1.2 release is used (649 songs).[4]

Of the annotations, only the structural annotations are retained. The structural annotations in the dataset follow the format and instructions established in the *SALAMI* project [182]. The transcriptions contain start and end times for every section and section labels for almost all sections. The section labels the annotators were allowed to assign were restricted to a list of 22, some of which were not used. The most frequently recurring section labels are: verse (34% of total annotated

3 E.g., by allowing for duplicates to give popular songs more weight, and by considering only chart notations up to 1991, to avoid some of the inconsistencies in how the Billboard charts themselves were compiled.

4 `http://ddmal.music.mcgill.ca/billboard`

time), chorus (24%), intro, solo, outro and bridge. The total number of sections, including the unlabeled ones, is 7762.

Are the annotations as reliable as the sample? Here we should note that there could be a hint of bias. The annotators guide, the instructions the annotators received, defines: "chorus (aka refrain): in a song, a part which contrasts with the verse and which is repeated more strictly".[5] The emphasis on the more 'strict' repetition in a chorus may skew the set of cues used by the annotators to the perform the section labeling task, towards repetition–related information.

4.2.2 *Audio Features*

A corpus analysis–centered study requires different kinds of descriptors than traditionally used in machine-learning applications. The descriptors are therefore selected based on the constraints put forward in chapter 3: we would like a set of features that is robust, informative and limited in size. Limiting the set to a small number of hand-picked descriptors is especially important since, for a part of the analysis, the amount of data required grows exponentially with the number of variables. The following is a list of the features selected for this analysis, beginning with the features computed for the *Billboard* dataset. Many of the features appear in the feature overview in chapter 2, so we will focus the discussion on their implementation. All features are one-dimensional.

Psycho-acoustic Features & Timbre

Loudness

The loudness descriptor is the standard psychological analogy of energy. It is obtained through comparison of stimuli spectra and a standardized set of equal loudness curves. We use the implementation by Pampalk [142]. The model applies outer-ear filtering and a spreading function before computing specific loudness values (N_k in sones) per

5 See `http://www.music.mcgill.ca/~jordan/salami/SALAMI-Annotator-Guide.pdf`

Bark band k and summing these values over all bands to obtain the total loudness T:

$$S = \max_k(N_k) + 0.15 \cdot \sum_{k \neq \max} N_k \tag{22}$$

where the factor 0.15 serves as a weighting that emphasizes the contribution of the strongest band. For every section, the loudness *mean* is computed and stored, as well as the inter-quartile range (*Loudness IQR*), as a measure of the section dynamics.

Sharpness
The sharpness descriptor is the psychoacoustic analog of the spectral centroid. It characterizes the balance between higher- and lower-band loudness. We use the Bark-specific loudnesses N_k as computed by Pampalk [142] and summing as formulated by Peeters [154]:

$$A = 0.11 \times \sum_k g(k) \cdot k \cdot N_k, \quad \text{where} \tag{23}$$

$$g(k) = \begin{cases} 1 & k < 15 \\ 0.066 \times \exp(0.171 \cdot k) & k > 15 \end{cases} \tag{24}$$

For every section, we use the *mean* sharpness. Compared to loudness range, sharpness range has no direct informative psycho-acoustic interpretation, so it is not included.

Roughness
Like the loudness descriptor, roughness is a mathematically defined psychoacoustic measure. It characterizes a timbral property of complex tones, relating to the proximity of its constituent partials. We use the MIRToolbox implementation by Lartillot et al. [103], which is based on a model by Plomp and Levelt [159]. Since the roughness feature has a very skewed distribution, it is summarized for every section by taking its *median*.

MFCC
As discussed in section 2.1.2, MFCC's are established multidimensional spectral envelope descriptors, designed to be maximally inde-

pendent. Individual MFCC coefficients tend to have no particular interpretation. In this model, therefore, the descriptor of interest is the variety in timbre. This is modeled by computing the trace of the square root of the MFCC covariance matrix, a measure of the timbre *total variance*. The MFCCs are computed following [142], and the first component (directly proportional to energy) is discarded.

Pitch Features

Chroma variance

Chroma features, also discussed in chapter 2.1.3, are widely used to capture harmony and harmonic changes. In the most typical implementation, the chroma descriptor or pitch class profile consists of a 12-dimensional vector, each dimension quantifying the energy of one of the 12 equal-tempered pitch classes. These energies can be obtained in several ways. The *NNLS chroma* features distributed along with the *Billboard* dataset are used in this study.[6]

In this study, the variety in the section's harmony is measured. Chroma, unlike MFCC, isn't typically looked at as a vector in Euclidean space, but rather as a distribution (of energy over pitch classes). Estimating just the total variance, as done for MFCC, would neglect the normalization constraint on chroma vectors and the dependencies it entails between pitch classes. We therefore normalize the chroma features per frame and assume it is Dirichlet-distributed. With the normalized features p as a 12-dimensional random variable, we can estimate a Dirichlet distribution from all of the section's chroma observations.

The 12-dimensional Dirichlet distribution $\mathcal{D}_{12}(\alpha)$, can be written:

$$f(p) \sim \mathcal{D}_{12}(\alpha) = \frac{\Gamma(\sum_{k=1}^{12} \alpha(k))}{\prod_{k=1}^{12} \Gamma(\alpha(k))} \prod_{k=1}^{12} p(k)^{\alpha(k)-1}, \tag{25}$$

6 http://www.isophonics.net/nnls-chroma

where Γ is the Gamma function. $\mathcal{D}_{12}(\alpha)$ can be seen as a distribution over distributions. We use the sum of the parameter vector $\alpha(k)$, commonly referred to as the Dirichlet precision s:

$$s = \sum_{k=1}^{12} \alpha(k) \tag{26}$$

It quantifies the difference between observing the same combination of pitches throughout the whole section (high precision) and observing many different distributions (low precision) [21]. There is no closed-form formula for s or α, but several iterative methods exist that can be applied to obtain a maximum-likelihood estimation (e.g., Newton iteration). Fast fitting was done using the *fastfit* Matlab toolbox by Minka.[7]

Pitch salience
The notion of pitch salience exists in several contexts. Here, it refers to the strength or energy of a pitch, specifically, the combined strength of a frequency and its harmonics, as in [168]. The *mean* of the strongest (per frame) pitch strength will be computed for every section.

Pitch centroid
As a last pitch-related feature, we include a notion of absolute pitch height, which is easy if the audio lends itself to reliable melodic pitch estimation. For the polyphonic pop music of the *Billboard* dataset, melody estimation is prone to octave errors. We therefore approximate pitch height in another way, using the more robust *Pitch centroid*. We define this as the average pitch height of all present pitches, weighted by their salience. Note that the pitch salience profile used here spans multiple octaves and involves spectral whitening, spectral peak detection and harmonic weighting in order to capture only tonal energy and emphasize the harmonic components. Our feature set includes the section *mean* of the pitch centroid as well as the interquartile range.

7 `http://research.microsoft.com/en-us/um/people/minka/software/fastfit/`

Melody Features

Contrary to the *Billboard* dataset, the *Dutch50* dataset contains songs with a mostly prominent melody, and relatively little post-processing. Specifically, it is less affected by the kind of heavy dynamic range compression that is commonplace in more recent popular music. This allows for a reasonable reliable melody estimate to be extracted for those songs (see the melody extraction challenges listed in section 2.1.4). The following features will therefore only be used for the *Dutch50* corpus analysis.

For all songs in the *Dutch50* dataset, the melody is extracted using the *Melodia* Vamp plug-in (see section 2.1.4, [168]). The resulting pitch contours and pitch salience are segmented along the annotated boundaries. For each section, statistics on the contour are then computed and compared.

Pitch strength
The melodic pitch strength, also referred to as pitch salience or salience function, is a measure of the strength of the fundamental frequency of the melody and its harmonics. For each section, the *mean pitch strength* was computed and normalized by subtracting the average pitch strength for the complete song.

Pitch height
For each section, the *mean pitch height* is computed and again normalized. A measure of pitch range was computed as well, in this case, the *standard deviation* of the pitch height.

Pitch direction
Finally, the *pitch direction* is estimated. With this measure, we aim to capture whether the pitch contours in a section follow an up- or downward movement. It is computed simply as the difference between the pitch height of the section's last and first half.

Structure Descriptors

Section length
The length of the section in seconds.

Section position
The position of the section inside the song is included as a number between 0 (the beginning) and 1 (the end).

4.3 CHORUSES IN EARLY POPULAR MUSIC

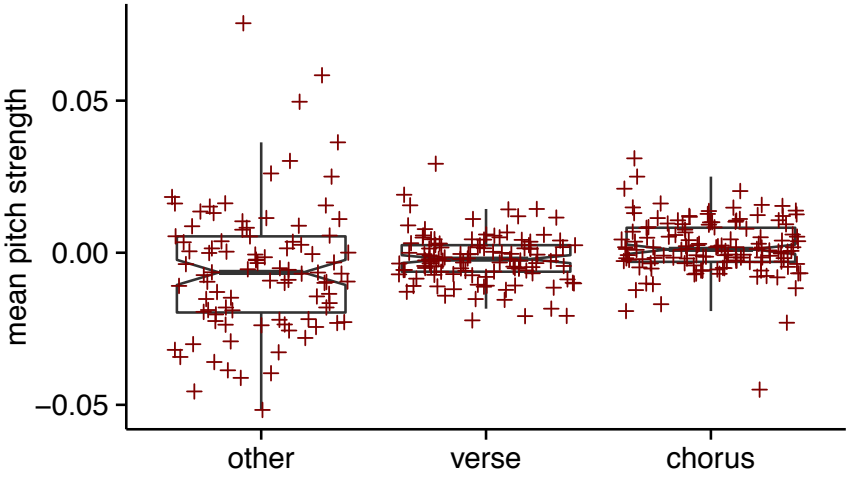
For the *Dutch50* dataset, we are more interested in melody than instrumentation and production, so in this first look at the features that make a chorus, the focus will be on melody.

The analysis will follow a simple approach, comparing raw feature differences between section types. This is easily manageable because of the constrained set of section labels. Nine songs were not considered as they contained only one type of section, in which case the labeling (verse or chorus or other) was found to be rather arbitrary. The remaining 41 songs contained a total of 330 sections, that were used to produce figures 18a – 19b.

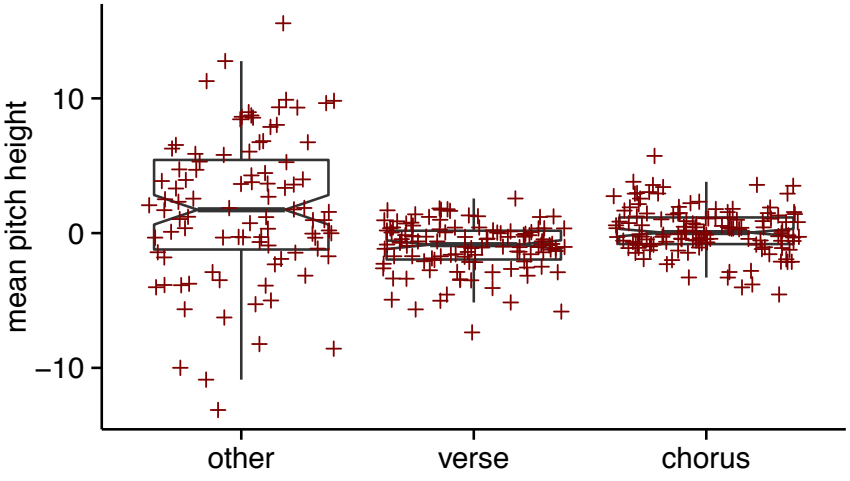
Figure 18a shows a scatter plot of the mean pitch strength values of all sections, over a box plot with estimates of the main quantiles (25%, 50% and 75%). A 95% confidence interval for the median is indicated by the notch in the box plot's sides. Remember that the mean pitch strength values for each section were normalized by subtracting the mean pitch strength over the complete song. The figure therefore illustrates how chorus pitch strengths do not significantly exceed the song average at 0, however, they are demonstrably higher than in verses and other sections. The former is confirmed in a t-test ($p < 0.001$) at a significance level, for this set of experiments, of 0.002.[8]

Figure 18b shows the *mean pitch height* for all sections, normalized by subtracting the overall song average. Correcting for multiple com-

8 We correct for multiple comparisons based on a total of 23 comparisons: 12 box plots, and 11 tests. $\alpha_{23} = \frac{0.05}{23} = 0.002$.

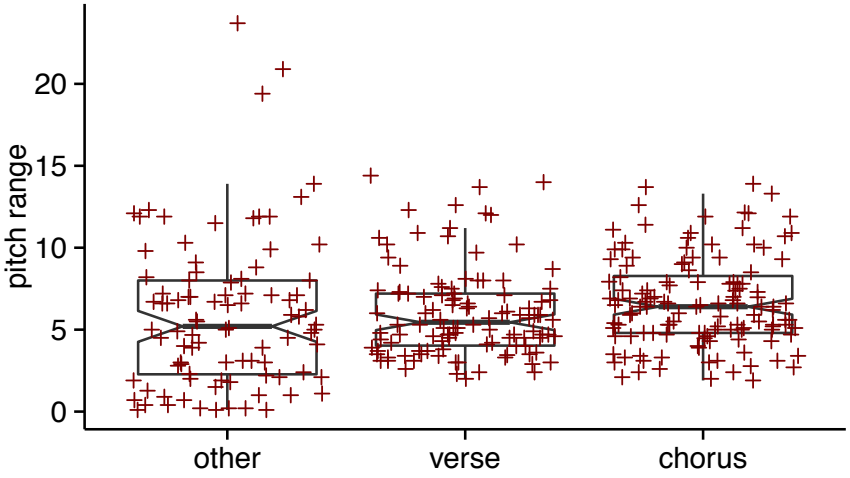


(a) Average pitch strength

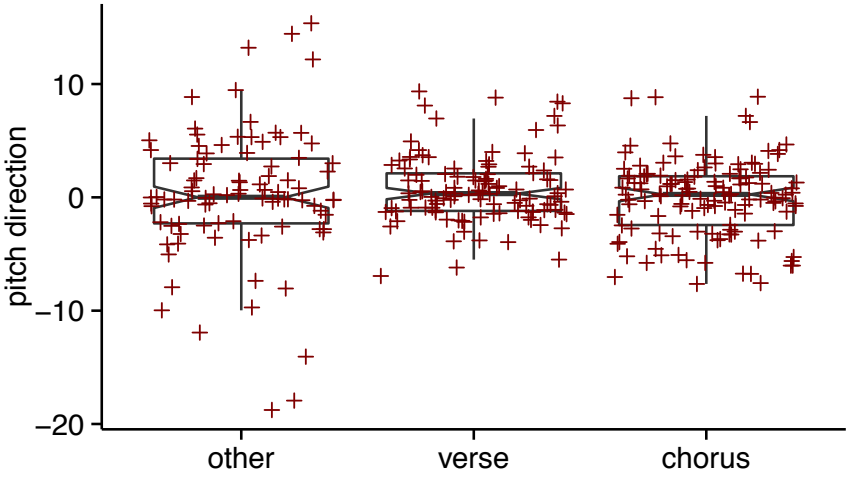


(b) Average pitch height

Figure 18.: Pitch statistics per section type in the *Dutch50* dataset.



(a) Pitch range



(b) Pitch direction

Figure 19.: Pitch statistics per section type in the *Dutch50* dataset.

parisons, pitch in the chorus is not significantly different from the song average ($p = 0.030$), but it is higher then the pitch of the verse, with over a semitone difference in median and mean ($p < 10^{-6}$ in a two-sided t-test). It is not significantly different from the pitch of the bridge and other sections ($p = 0.004$, two-sided Welch's t-test).

Figures 19a and 19b show the pitch range and direction for all sections (not normalized). Pitch ranges are wider for choruses than for verses and other sections, but not significantly. Finally, average pitch direction shows no trend for choruses at all. The average direction for verses is greater than zero with $p = 0.003$, suggesting an upward tendency in pitch during the verse, but again, this is not significant when significance level is adjusted for multiple comparisons, so that no conclusions can be made from this observation.

Summing up the findings, the analysis shows how choruses in the *Dutch50* dataset have a stronger and higher pitch than verses. Note that several more trends would have emerged from these test statistics if the confidence level hadn't been corrected for. The correction is nonetheless crucial: the initial exploration using box plots, as well as the subsequent tests must be accounted for, as argued in Chapter 3.

4.4 CHORUSES IN THE BILLBOARD DATASET

For the *Billboard* analysis, all descriptors are used except for those pertaining to melody, since melody estimates are expected to the substantially less accurate then they were for the *Dutch50* data. The resulting features make up a dataset of 7762 observations (sections) and 12 variables (descriptors) for each observation: the above perceptual features and one section label. These data will be used to model what features correlate with a section being a chorus or not.

More specifically, they will be modeled using a PGM, or probabilistic graphical model. We now explain the concept of Probabilistic Graphical Models (PGM) by introducing three varieties of graphical models: correlation graphs, partial correlation graphs, and Bayesian networks.

4.4.1 *Graphical Models*

Graphs and networks can be very useful to conceptualize the relations between random variables. The easiest model to display relations between variables is the *correlation graph*. It's a graph in which all variables are nodes, and two variables are connected by an edge if and only if they are correlated:

$$E_c(i,j) = \begin{cases} 1 & \left|\rho(X_i, X_j)\right| > \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

This is, the matrix E_c encoding the correlation graph's edges contains a 1 wherever the absolute correlation (e.g., Pearson correlation ρ) between the corresponding variables (e.g., X_i and X_j) is greater then some threshold ε.

Correlation graphs are useful as visualizations, but can be quite dense, with correlations between many of the variables. They also provide little information about the underlying multivariate distribution of $(X_1, \ldots, X_n)$.

A more widely used type of graphs are *partial correlation graphs*. The partial correlation between two variables X_i and X_j is the correlation between ϵ_i and ϵ_j: the errors for X_i and X_j after removing the effects of all other variables $\{X_k\}$. Specifically, the correlation between the errors of X_i and X_j after linear regression with $\{X_k\}$. This sheds some more light on each variables' contribution to the dependencies among the set of variables than a simple correlation graph.

$$E_{pc}(i,j) = \begin{cases} 1 & \left|\rho(\epsilon_i, \epsilon_j)\right| > \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

Finally, the graphs that are most often referred to when talking about Probabilistic Graphical Models are *Bayesian networks*. Bayesian networks encode *conditional independence*. When two variables in a PGM are *not* connected, they are conditionally independent given the other variables:

$$E_{PGM}(i,j) = \begin{cases} 0 & X_i \perp\!\!\!\perp X_j \text{ given } X_k \ \forall k \\ 1 & \text{otherwise} \end{cases} \tag{29}$$

If all variables are normally distributed (the distribution is multivariate Gaussian), and the graph is undirected, E_{PGM} will be the same as the partial correlation graph E_{pc} [200]. Generally, however, Bayesian networks contain directed edges (i.e., arrows), encoding a special kind of dependence in which some variables are *parents* of other variables. Bayesian networks are also acyclic, i.e., they contain no cycles. This allows us to identify for each variable, not just its parents and children, but also its *ancestors* and *descendants*. In a fully directed Bayesian network, a more specific independence property holds: any variable X_j is conditionally independent of its non-descendants, given its parents [21]. The latter implies a particular relationships between the conditional distributions of the variables and their joint distribution $p(X_1, \ldots, X_n)$:

$$p(X_1, \ldots, X_n) = \prod_{k=1}^{n} p(X_k | pa_k) \tag{30}$$

where pa_k denotes the set of parents of X_k. In other words, by defining pa_k, directed graphical models straightforwardly encode a factorization of the joint probability distribution $\mathrm{p}(X_k)$ underlying the graph.

Graph structures of Bayesian networks are typically constructed using prior expert knowledge, but they can also be learned from data. However, when considering a set of variables in the real world, it will not usually be possible to know the direction of every edge. When learning a PGM's graph structure from data, even if all conditional dependence relations are known, one set of conditional independences can often be represented by several Bayesian networks, with arrows pointing in different directions. For this and a number of other reasons, it is dangerous to interpret edge directions in Bayesian networks as *causal* relationships.

Yet, directed and partially directed graphical models can nonetheless be interesting: the absence of edges does encode independence with other variables controlled for, and sometimes, some edge directions may indeed be found, in which case an interpretation of the edge directions can help to assess whether the model makes sense. For more on the interpretation of Bayesian networks and examples for music analysis, see [21].

PGM Structure Learning

Learning the PGM structure generally requires a great amount of conditional independence tests. The *PC algorithm* optimizes this procedure and, in addition, provides information about the direction of the dependencies where they can be inferred [86]. When not all directions are found, a partially directed graph is returned.

One of the limitations of the PC algorithm, however, is that the variables must be either all discrete, or all continuous, following a normal distribution. In the analysis in the next section, all data are modeled as continuous. For most variables, this is straightforward, except for the *Section Type* variable, which will have to be remodeled as continuous. We do this by introducing the notion of *Chorusness*.

4.4.2 *Chorusness*

The Chorusness variable is derived from the *Chorus probability* p_C, a function over the domain of possible section labels. The chorus probability $p_C(T)$ of a section label T is defined as the probability of a section annotated with label T being labeled 'chorus' by a second annotator. In terms of the annotations T_1 and T_2 of two independent and unbiased, but 'noisy' annotators, $p_C(T)$ can be written:

$$p_C(T) = p(T_1 = C|T_2 = T) = p(T_2 = C|T_1 = T), \tag{31}$$

where C refers to the label 'chorus'.

The *Billboard* dataset has been annotated by only one expert per song, therefore it contains no information about any of the $p_C(T)$. However, in the *SALAMI* dataset, annotated under the same guidelines and conditions, two independent annotators were consulted per song [182]. The annotators' behaviour can therefore be modeled by means of a confusion matrix $M(T_1, T_2) \in [0, 1]^{22\times22}$ between all 22 section types:

$$M(T_1, T_2) = f(x_1 = T_1 \cap x_2 = T_2) \tag{32}$$

with frequencies f in seconds (of observed overlapping labels T_1 and T_2). Since the identities of the two annotators have been randomized, M may be averaged out to obtain a symmetric confusion matrix $M^\star$:

$$M^\star = \frac{M + M^T}{2} \tag{33}$$

From here we can obtain the empirical Chorus probability:

$$p_C(T) = \frac{M^\star(T, C)}{\sum_k M^\star(T, k)} \in [0, 1]. \tag{34}$$

Chorus probability values for every section type were obtained from the *Codaich-Pop* subset of the *SALAMI* dataset (99 songs). Finally, the Chorus Probability is scaled monotonically to obtain the Chorusness measure $C(T)$, a standard *log odds ratio* of p_C:

$$C(T) = \log\left(\frac{p_C(T)}{1 - p_C(T)}\right) \in (-\infty, \infty). \tag{35}$$

It ranges from −8.41 (for the label 'spoken') to 0.83 (for the label 'chorus').

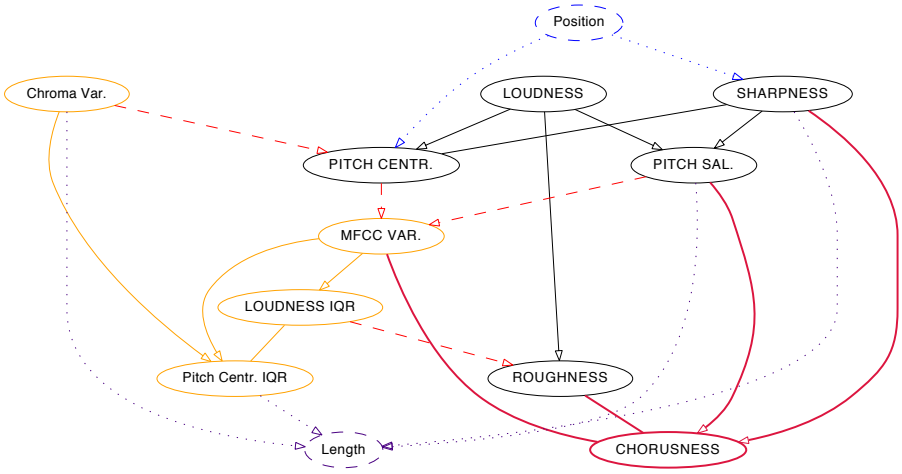
4.4.3 *Implementation*

Before the model learning, a set of Box-Cox tests is performed to check for rank-preserving transformations that would make any of the variables more normal. The Chroma variance s is found to improve with a power parameter $\lambda = -1$, and therefore scaled as:

$$S = \frac{s^\lambda - 1}{\lambda} = 1 - \frac{1}{s} \tag{36}$$

The Section length, Loudness IQR and Pitch centroid IQR are found to improve with a log transform. Weeding out divergent entries in the dataset leaves us with a subset of 6462 sections and 12 variables.

The R-package *pcalg* implements the PC-algorithm. Beginning with a fully connected graph, it estimates the graph skeleton by visiting all pairs of adjacent nodes and testing for conditional independence



Note. Bold edges highlight the features that correlate with Chorusness. Black edges are edges between features that are closely related on the signal level. The orange, lighter edges denote relations between features that represent some kind of variance. Dotted edges are used for the features that are not measured from the audio (Position and Length). The remaining edges are drawn as dashed, red lines.

Figure 20.: Graphical model of the 11 analyzed perceptual features and Chorusness variable C. $\alpha_{\text{PGM}} = 0.05$. The edges are colored by the author to facilitate discussion

given all possible subsets of the remaining graph.[9] The procedure is applied to the 6462×12 dataset, with 'conservative' estimation of directionality, i.e. no direction is forced onto the edges when the algorithm cannot estimate them from the undirected graph structure.

4.4.4 *Analysis Results*

The resulting graphical model is shown in Figure 20. It is obtained with $p < 3.5 \times 10^{-5}$, the significance level required to bring the over-

9 `http://cran.r-project.org/web/packages/pcalg/`

all probability of observing one or more edges due to chance, under 5 percent. In terms of the significance level α_{CI} of the conditional independence tests and α_{PGM} of the model:

$$\alpha_{\text{CI}} = 1 - (1 - \alpha_{\text{PGM}})^{1/n} \approx \frac{\alpha_{\text{PGM}}}{n} \tag{37}$$

with $\alpha_{\text{PGM}} \ll 1$ (here 0.05) and n the number of tests performed ($\sim$ 1500).

Note that $p \approx 10^{-5}$ is a conservative parameter setting for an individual test. As a result, it is best to view the model as a depiction of dependencies rather than independences, since the latter may always be present at a lower significance than required by the α.

The model is relatively stable with respect to α: it is unchanged when the learning procedure is repeated with more restrictive $\alpha_{\text{PGM}} = 0.01$ and 0.005. Four additional edges appear with a more tolerant $\alpha = 0.10$.

4.4.5 *Discussion*

A number of observations can be made about the chorusness model in figure 20. First, the most expected dependencies in the model are highlighted.

At least three kinds of feature relations are expected. First, there are the correlations between features that are closely related on the signal level (black edges): Loudness and Pitch salience, for example, measure roughly the same aspects of a spectrum (and can be expected to be proportional to roughness), and so do Sharpness and Pitch centroid. Roughness is a highly non-linear feature that is known to be proportional to energy. The model reflects this.

The second kind of correlations are the relations between variance-based features and the Section length variable. Musically, it is expected that longer sections allow more room for an artist to explore a variety of timbres and pitches. This effect is observed for Chroma variance and Pitch centroid IQR, though not for MFCC variance and Loudness IQR. Interestingly, correlations with Section length point *towards* it rather than away (dotted edges): the length of a section length

is a result of its variety in pitch and timbre content, rather than a cause. Note again, however, that directions of effects in a learned PGM are not always reliable enough to be taken at face value.

Third, some sections might just display more overall variety, regardless of the section length. This would cause different variances to relate, resulting in a set of arrows between the four variance features. Four such relations are observed (lighter, orange edges).

We now note that Sharpness, Pitch salience and Roughness predict Chorusness, as well as the MFCC variance (bold edges). All of these can be categorized as primarily timbre-related descriptors. Section length, Section position and Chroma variance are d-separated from Chorusness, i.e., no direct influence between them has been found. The status of Pitch centroid, Loudness, and Loudness IQR is uncertain. Depending on the true direction of the Chorusness, MFCC variance and Roughness relations, they may be part of the Chorusness *Markov blanket*, the set of Chorusness' parents, children, and parents of children, which d-separates Chorusness from all other variables, or they might be d-separated themselves (given the Markov blanket, no influence between these variables and Chorusness) [93].

Also interesting are the more unexpected dependencies. For example, two variables depend directly on the Section position, while Chorusness does not. This may be due to the limitations of the normal distribution by which all variables are modeled; it might not reflect the potentially complex relation of Chorusness variable and Section position. However, the Section position variable does predict Sharpness and Pitch centroid to some extent (dotted edges). A simple regression also shows both variables correlate *positively*, suggesting an increased presence of sharper and higher-pitched sections towards the end of the songs in the *Billboard* corpus.

Finally, the dashed red edges in the diagram indicate dependencies that are most unintuitive. Tentative explanations may be found, but since they have no effect on Chorusness, we will omit such speculations here.

		95% CI	
	β	LL	UL
Sharpness	0.11	0.10	0.13
MFCC variance	0.12	0.09	0.15
Roughness	0.12	0.08	0.16
Pitch salience (×10)	0.04	0.03	0.05
Loudness	0.03	-0.01	0.06
Loudness IQR	-0.33	-0.48	-0.18
Pitch centroid	0.10	0.07	0.12

Table 1.: Results of a multivariate linear regression on the Chorusness' Markov blanket.
CI=confidence interval, LL=lower limit, UL=upper limit.

4.4.6 *Regression*

We ran a regression model to see in more detail how the set of features related to Chorusness predict our variable of interest. Table 1 lists the coefficients of a linear regression on the Chorusness variable and its Markov blanket, i.e. those variables for which a direct dependency with Chorusness is apparent from the model. Since there is no certainty about the exact composition of the Markov blanket, all candidates are included, including Loudness, Loudness IQR and Pitch centroid. Note that, having defined Chorusness as a log odds ratio, this linear regression is in effect a logistic regression on the section's original Chorus probability $p_C \in [0, 1]$.

One can see that all features but the Loudness IQR have positive coefficients. Only loudness has no significant positive or negative correlation. We conclude that, in this model, sections with high Chorusness are sharper and rougher than other sections. Chorus-like sections also feature a slightly higher and more salient pitch, a smaller dynamic range and greater variety in MFCC timbre.

4.4.7 *Validation*

Finally, a classification experiment is performed. It consists of the evaluation of a 2-way classifier that aims to label sections as either 'chorus' or 'non-chorus'. A k-nearest neighbor classifier ($k = 1$) is trained on half of the available sections, and tested on the other half (randomly partitioned). This procedure is repeated 10 times to obtain an average precision, recall and F-measure.

The results confirm the trends found in the PGM: using just the Markov blanket features of table 1, the classifier performs better than random: $F = 0.52$, 95% CI $[0.51, 0.52]$ vs. a maximum random baseline of $F = 0.36$. The classifier also performs better than one that uses all features ($F = 0.48$), or only Loudness and Loudness IQR ($F = 0.48$), the features used in [148].

4.5 CONCLUSIONS

This chapter presents two computational studies into the robust and informative audio descriptors that correlate with the 'chorusness' of sections in pop songs. A selection of existing and novel perceptual and computational features is presented, and applied to two datasets. A small new dataset of early Dutch popular songs, *Dutch50*, is analyzed to reveal that choruses in the dataset have stronger and higher pitch then verses. A larger dataset, the *Billboard* dataset, has been analyzed using a probabilistic graphical model and a measure of Chorusness that is derived from annotations and an inter-annotator confusion matrix. The resulting model was complemented with a regression on the most important variables. The results show that choruses and chorus-like sections are sharper and rougher and, like the pre-1950 Dutch choruses, feature a higher and more salient pitch. They have a smaller dynamic range and greater variety of MFCC-measurable timbre than other sections. These conclusions demonstrate that, despite the challenges of audio corpus analysis presented in the previous chapter, musical insights can be gained from the analyses of readily

available datasets. Moreover, having been guided by the desiderata in the previous section, we believe our insights are robust.

The results obtained in a classification experiment do not suggest that our model would read the level of accuracy obtained by the state-of-the-art techniques that incorporate repetition information. However, they demonstrate for the first time that there is a class of complementary musical information that, independently of repetition, can be used to locate choruses. This suggests that our model can be applied to complement existing structure analysis applications, while repetition information and section order can in turn enhance our model of Chorusness for further application in popular music cognition research and audio corpus analysis.

5

COGNITION-INFORMED PITCH DESCRIPTION

5.1 INTRODUCTION

In empirical, corpus-based music research, we may want to be able to describe high-level, cognition-related qualities of music, such as its complexity, expectedness and repetitiveness, from raw audio data. The features we would need to do this have not gotten the attention they deserve in MIR's audio community, perhaps due to the 'success' of low-level features when perceptual and cognitive validity is not a concern (i.e., most of the time—see chapter 3).

In this chapter, we propose a set of novel cognition-informed and interpretable descriptors for use in mid- to high-level analysis of pitch use, and applications in content-based retrieval. They are based on symbolic representations that were originally inspired by results in music cognition, and have been since been shown to work well in symbolic music analysis. We focus on features that describe the use of *pitch*, which we use here not just in its perceptual definition (a psycho-acoustic dimension related to the frequency of a sound event), but in a wider sense that includes both *harmony* and *melody*. In the long run, we believe, better mid-level and high-level pitch descriptors will provide insight into the building blocks of music, including riffs, motives and choruses.

In the second part of this chapter, we test the new descriptors in a cover song detection experiment. At the end of this section, we motivate why this type of application can serve as a good test case. Sections 5.3.1 – 5.3.3 present the data, methods, and results.

5.1.1 *Improving Pitch Description*

Symbolic Music Description and Music Cognition

In chapter 1.2.4, we discussed the theoretical arguments for symbolic vs. audio-based music representations. Both representations also have *practical* advantages. Symbolic music representations encode music in terms of discrete events. Discrete music events such as notes can be *counted*, making statistical modeling more straightforward (see many of the systems reviewed by Burgoyne in [21]). Symbolic representations also allow more easily for models that acknowledge the order of events in time or look for hierarchical structure on the music (see de Haas [61] and chapter 3). None of these abstractions are easily accessed through currently available audio features. This may explain why symbolic music has been the representation of choice in all of the corpus-based music cognition research reviewed in chapter 3. Since we aim to develop new audio representations that get a step closer to describing cognition-level qualities of music, we may be able to take some inspiration from the technologies that exist in symbolic music description.

Rhythm description is not included in this chapter. The state of the art in rhythm description, e.g., measuring inter onset intervals or syncopation, requires a reliable method of estimating and characterizing streams of salient onsets. Robust rhythm description thus involves two of the most difficult remaining challenges in MIR right now: note segmentation in a polyphonic mix, and separation of notes into streams (see section 2.1.4 on melody extraction). This makes rhythm description on any level beyond tempo and meter very difficult with the current state of the art in the above tasks, similarly to how complete music transcription isn't robust enough to be useful at present time. Timbre is also not considered in this chapter, as it is not typically described in symbolic terms as much as melody, harmony and rhythm. We come back to this point in the conclusions in chapter 9.

We focus on harmony and melody, or 'pitch'. Melodic pitch estimation, discussed in section 2.1.4, involves fewer steps than music transcription, making it quite reliable in comparison. To ensure robustness, however, we will make two further restrictions on the kind of representations we pursue.

1. melody description should not require note segmentation
2. harmony description should not involve chord estimation

The state-of-the-art in chord estimation is considerably more successful then rhythm transcription, yet any transcription in general runs the risk of introducing unknown biases due to the assumptions of the transcription systems (e.g., for many chord transcriptions: datasets on which it was trained or evaluated). See also section 3.2.3 for a more elaborate discussion of this issue.

Finally, we add two more restrictions.

3. descriptors should be invariant to non-pitch-related facets such as timbre, rhythm and tempo
4. descriptors should have a fixed size

Chroma features or pitch class profiles are a proven and relatively robust representation of harmony but, like most feature time series, vary in length with the audio from which they have been extracted, and are not tempo- and translation-invariant. An adequate fixed-size descriptor should capture more detail than a simple chroma pitch histogram, while preserving tempo and translation invariance.

5.1.2 *Audio Description and Cover Song Detection*

In the second part of this chapter, three new descriptors will be evaluated. We now argue that the task of *scalable cover song retrieval* is very suitable for developing descriptors that effectively capture mid- to high-level musical structures, such as chords, riffs and hooks.

Cover detection systems, as explained in section 2.2.3, take a query song and a database and aim to find other versions of the query song.

Most successful cover detection algorithms are built around a two-stage architecture. In the first stage, the system computes a time series representation of the harmony or pitch for each of the songs in a database. In the second stage, the time series representing the query is compared to each of these representations, typically through some kind of alignment, i.e., computing the locations of maximum local correspondence between the two documents being compared. Such alignment methods are very effective, but computationally expensive.

Consider an archivist or musicologist, who aims to exhaustively search for musical relations between documents in a large audio corpus, e.g., an archive of folk music recordings. Archives like this may contain large numbers of closely related documents, such as exact duplicates of a recording, different renditions of a song, or variations on a common theme. The particularities of such variations are of great interest in the study of music genealogies, oral transmission of music, and other aspects of music studies [196].

Cover song detection can be helpful here, but alignment-based techniques are no longer an option: a full pair-wise comparison of 10,000 documents could easily take months.[1] This is why some researchers have been developing the more scalable cover song detection techniques reviewed in section 2.2.3. Scalable strategies are often inspired by audio fingerprinting and involve the computation of an indexable digest of (a set of) potentially stable landmarks in the time series, which can be stored and matched through just a few inexpensive look-ups.

The challenges laid out above make cover song detection an ideal test case to evaluate a special class of descriptors: harmony, melody and rhythm descriptors, global or local, which have a fixed dimensionality and some tolerance to deviations in key, tempo and global structure. If a collection of descriptors can be designed that accurately describes a song's melody, harmony and rhythm in a way that is ro-

1 MIREX 2008 (the last to report run times) saw times of around 1.2–3.2 seconds per comparison. These algorithms would take 1.8–6 years to compute the $\frac{1}{2}10^8$ comparisons required in the above scenario, or 6–20 weeks at current processor speeds (assuming eight years of Moore's law—processor speeds doubling every 2 years).

bust to the song's precise structure, tempo and key, we should have a way to determine similarity between the musical 'gist' of two songs and assess if the underlying composition is likely to be the same.

Note that the interest, in this case, is not necessarily in the large-scale performance or efficiency of the system, but in the evaluation of a fixed-sized descriptor in the context of performance variations.

5.2 COGNITION-INSPIRED PITCH DESCRIPTION

Symbolic Music Description and Expectation

The possibility of computing statistics on discrete musical events, and the possibility of representing time, have inspired some researchers to use music data to test models of musical expectation. There is an increasing amount of evidence that the primary mechanism governing musical expectations is statistical learning [78, 153]. On a general level, this implies that the relative frequencies of musical events play a large role in their cognitive processing. Expectations resulting from the exposure to statistical patterns have been shown to affect the perception of melodic complexity and familiarity, preference, and recall [78], making them a particularly interesting for some of the applications in this thesis.

Statistical distributions of musical events are often modeled using the notion of *bigrams*. Bigrams are, simply put, ordered pairs of observations. Word bigrams and letter bigrams, for example, are much-used representations in natural language processing and computational linguistics [116]. Figure 21 shows the set of word bigrams extracted from the phrase "to be or not to be".

In *melody* description, numerous authors have proposed representations based on pitch bigrams, most of them from the domain of cognitive science [110, 134, 166]. This comes as no surprise: distributions of bigrams effectively encode two types of probabilities that influence expectation: the prior probability of pairs of pitches, and, if we condition on the first pitch in each pair, the conditional frequency of a pitch given the one before.

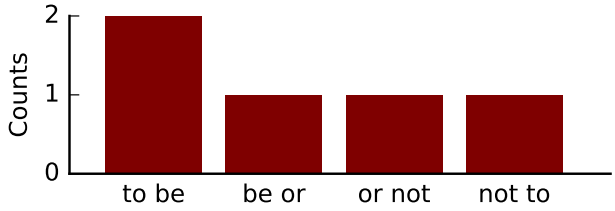


Figure 21.: A bigram count for the phrase "to be or not to be".

In the description of *harmony*, bigram and other n-gram representations have been used as well. In [167], for example, Rohrmeier and Graepel find that chord bigram-based models predict chord sequences better than plain HMM on a subset of the *Band-in-a-box* corpus of Jazz chord sequences, though not as good as trigrams, higher-order n-grams and autoregressive HMM.

When symbolic data are not available, bigrams in the strict sense are difficult to compute. For this reason, the features we propose are approximations of the notion of bigrams that can also be interpreted as probability distributions, encode a notion of order in time, and can be computed from audio. The descriptors will be named *audio bigrams*.

5.2.1 *Pitch-based Audio Bigrams*

First, we propose three pitch bigram-like representations.

The Pitch Bihistogram

The first new feature is a melody descriptor. It essentially captures how often two pitches p_1 and p_2 occur less than a distance d apart.

Consider a 12-dimensional melody time series $M(t, p)$. As in chroma, M contains pitch activations, quantized to semitones and folded to one octave. If a pitch histogram is defined as:

$$h(p) = \sum_t M(t, p), \tag{38}$$

with $p \in \{1, 2, \ldots 12\}$, the proposed feature is then defined:

$$PB(p_1, p_2) = \sum_t M(t, p_1) \max_\tau (M(t + \tau, p_2)) \tag{39}$$

with $\tau = 1, 2 \ldots \Delta t$. This will be referred to as the *pitch bihistogram* (*PB*), a bigram representation that can be computed from any melodic pitch time series, up to arbitrarily high frame rates. Note that the use of pitch classes rather than pitch creates an inherent robustness to octave errors in the melody estimation step, making the feature insensitive to one of the most common errors encountered in pitch extraction.

Alternatively, scale degrees can be used instead of absolute pitch class. In this scenario, the melody time series $M(t, p)$ must first be aligned to an estimate of the piece's overall tonal center. As a tonal center, the tonic can be used. However, for extra robustness to mis-estimation of the tonic, we suggest to use the tonic for major keys and the minor third for minor keys, so that mistaking a key for its relative major or minor has no effect.

Chroma Correlation Coefficients

The second feature representation we propose looks at vertical rather than horizontal pitch relations. It encodes which pitches appear simultaneously in a 12-dimensional chroma time series $H(t, p)$. From $H(t, p)$ we compute the correlation coefficients between each pair of chroma dimensions to obtain a 12×12 matrix of *chroma correlation* coefficients $CC(p_1, p_2)$:

$$CC(p_1, p_2) = \sum_t H^\star(t, p_1)\ H^\star(t, p_2), \tag{40}$$

in which $H^\star(t, p)$ is $H(t, p)$ after column-wise normalization, i.e., subtracting, for every p, the mean and dividing by the standard deviation:

$$H^\star = \frac{H(t, p) - \mu(p)}{\sigma(p)} \tag{41}$$

$$\begin{aligned} \mu(p) &= \tfrac{1}{n} \textstyle\sum_t H(t, p) \\ \sigma^2(p) &= \tfrac{1}{n-1} \textstyle\sum_t (H(t, p) - \mu(p))^2 \end{aligned} \tag{42}$$

This descriptor is similar to the chroma covariance feature proposed by Kim in [90]. Like the pitch bihistogram, the chroma features can be transposed to the same tonal center (tonic or third) based on an estimate of the overall or local key.

Harmonization
Finally, the harmonization feature (*HA*) is a set of histograms of the harmonic pitches $p_h \in \{0, \ldots, 11\}$ as they accompany each melodic pitch $p_m \in \{0, \ldots, 11\}$. It is computed from the pitch contour $P(t)$ and a chroma time series $H(t, p_h)$, which should be adjusted to have the same sampling rate.

$$HA(p_m, p_h) = \sum_t M(t, p_m)\ H(t, p_h). \tag{43}$$

Musically, the harmonization feature summarises how each note of the pitch tends to be harmonised.

From a memory and statistical learning perspective, the chroma correlation coefficients and harmonization feature may be used to approximate expectations that include: the expected consonant pitches given a chord note, the expected harmony given a melodic pitch, and the expected melodic pitch given a chord note. Apart from [90], where a feature resembling the chroma correlation coefficients is proposed, information of this kind has yet to be exploited in a functioning (audio) MIR system. Like the pitch bihistogram and the chroma correlation coefficients, the harmonization feature has a dimensionality of 12×12.

5.2.2 *Pitch Interval-based Audio Bigrams*

The next three descriptors extend the pitch-based audio bigrams above to interval representations. Whereas pitch bigram profiles are expected to strongly correlate with the key of an audio fragment, interval bigrams are key-invariant, which allows them to be compared across songs.

Melodic Interval Bigrams

The melodic interval bigrams (*MIB*) descriptor is a two-dimensional matrix that measures which pairs of pitch intervals follow each other in the melody It is based on *pitch trigrams*, an extension of the two-dimensional bihistogram in equation 39:

$$\text{trigrams}(p_1, p_2, p_3) = \sum_t \max_\tau(M(t-\tau, p_1))\ m(t, p_2) \max_\tau(M(t+\tau, p_3)), \tag{44}$$

with again $\tau = 1 \ldots \Delta t$ and M the melody matrix, the binary chroma-like matrix containing the melodic pitch class activations. The result is a three-dimensional matrix indicating how often triplets of melodic pitches (p_1, p_2, p_3) occur less than Δt seconds apart.

The pitch class triplets in this feature can be converted to interval pairs using the function:

$$\text{intervals}(i_1, i_2) = \sum_{p=0}^{11} X((p - i_1) \bmod 12,\ i,\ (p + i_2) \bmod 12). \tag{45}$$

This maps each trigram (p_1, p_2, p_3) to a pair of intervals $(i_2 - i_1, i_3 - i_2)$. A broken major chord $(0, 4, 7)$ would be converted to $(4, 3)$, or a major third followed by a minor third. Applied to the pitch trigrams, the intervals function yields the *melodic interval bigrams* descriptor:

$$MIB(i_1, i_2) = \text{intervals}(\text{trigrams}(M(t, p))) \tag{46}$$

Harmonic Interval Co-occurrence

The harmonic interval co-occurrence descriptor measures the distribution of triads in an audio segment, represented by their interval representation. It is based on the *triad profile*, which is defined as the three-dimensional co-occurrence matrix of three identical copies of the chroma time series $H(t, p)$ (t is time, p is pitch class):

$$\text{triads}(p_1, p_2, p_3) = \sum_t H(t, p_1)\ H(t, p_2)\ H(t, p_3). \tag{47}$$

The triad profile can be made independent of absolute pitch by applying the intervals function (equation 45). This yields the *harmonic interval co-occurrence* matrix:

$$HIC(i_1, i_2) = \text{intervals}(\text{triads}(H(t, p))) \tag{48}$$

As an example, a piece of music with only minor chords will have a strong activation of HIC$(3, 4)$, while a piece with a lot of tritones will have activations in HIC$(0, 6)$ and HIC$(6, 0)$.

Harmonization Intervals
Finally, the harmonization feature can be extended to obtain the harmonization intervals (HI) feature, defined as:

$$HI(i) = \sum_t \sum_{p=0}^{12} M(t, p)\ H(t, (p + i) \bmod 12) \tag{49}$$

Unlike the 12×12 *MIB* and *HIC*, the *HI* is 12-dimensional, and measures the distribution of intervals between the melody and harmony.

5.2.3 *Summary*

We have proposed three 12×12 melody and harmony descriptors based on pitch, and two 12×12 and one 12-dimensional descriptor based on pitch intervals. The first three will now be used in an experiment to test how much harmonic and melodic information they encode. The last three features will be used in part iii of this thesis, where a key-invariant descriptor is needed.

Finally, we note that the term 'audio bigrams' assumes a necessarily loose interpretation of the term bigrams. It is a loose interpretation in that not all pitch pairs in the pitch bihistogram follow each other immediately—some other pitch content might be present in between—and pitch pairs in the chroma correlation feature are simultaneous rather than adjacent.[2] However, this loose interpretation is necessary if we want to apply the idea of bigrams to audio at all. Because of the

2 similar to 'skipgrams' in natural language processing.

continuous nature of audio signals, there is no notion of 'adjacency' without resorting to some arbitrary discretization of time—we can only measure what is 'close together'.[3] The use of the word bigrams will also become more clear as we define the concept of audio bigrams more formally in the next chapter.

5.3 EXPERIMENTS

To test whether the features we propose capture useful information, we perform a number of cover song detection experiments.

5.3.1 *Data*

Two datasets are used: *covers80*, a standard dataset often used as as a benchmark, and the *translations* dataset. The *covers80* dataset is a collection of 80 cover song pairs, divided into a fixed list of 80 queries and 80 candidates. Results for this dataset have been reported by at least four authors [174], and its associated audio data are freely available. It is not as big as the much larger *Second Hand Song* dataset.[4] The problem with the Second Hand Song dataset, however, is that it is distributed only in the form of standard Echo Nest features. These features do not include any melody description, which is the basis for two of the descriptors proposed in this chapter.

The *translations* dataset is a set of 150 recordings digitized especially for this study: 100 45-rpm records from the 50's and 60's, and 50 78-rpm records, most of them from before 1950. They have been selected from the collections of the Netherlands Institute for Sound and Vision, who own a 'popular music heritage' collection that was, until recently,

3 Of course, frequency-domain representations of signals come in discrete frames by construction, but to choose this same discretization for measuring pitch transitions would restrict the scope of our features to pitch patterns on very short time scales. A discretization based on onsets or beats may be more meaningful. Here, however, we argue that the feature should be meaningful both in the presence and absence of rhythm.

4 `http://labrosa.ee.columbia.edu/millionsong/secondhand`

only accessible in the form of manually transcribed metadata (such as titles, artists, original title, composer).

Amongst these records are 50 pairs of songs that correspond to the same composition. All of these tunes are translated covers or re-interpretations of melodies with a different text—in the early decades of music recording, it was very common for successful singles to be re-recorded and released by artists across Europe, in their own language.[5] Such songs are especially interesting since they guarantee a range of deviations from the source, which is desirable when models of music similarity are tested. Some pairs are re-recordings by the original artists, and thus very similar, other song pairs need a very careful ear to be identified as 'the same'.

5.3.2 *Methods*

Features

Four experiments were carried out, each following the same retrieval paradigm. The features that will be used are the three audio pitch bigram representations proposed in section 5.2.1: the pitch bihistogram, chroma correlations coefficients and the harmonization feature.[6]

Similarity

A query song is taken out of the collection and its feature representation is computed. The representations for all candidate songs are then ranked by similarity, using the *cosine similarity* $s_{\cos}$,

$$s_{\cos}(x, y) = \cos(\alpha_{xy}) = \frac{x \cdot y}{\|x\| \, \|y\|}, \tag{50}$$

5 from correspondence with the curators at the Netherlands Institute for Sound and Vision and Meertens Institute

6 The other, pitch interval-based audio bigrams had not been formalized yet at the time when this experiment was carried out, but close variants of theses features were found to perform no better than the pitch-based audio bigrams in initial tests.

where α_{xy} is the angle between the vectors x and y. Note that, even though no indexing is used, cosine similarities can be computed much faster than alignment-based distances.

Evaluation

To evaluate the results, three evaluation measures are used. 'Recall at 1' (R_1) is the fraction of queries for which the correct song is returned in first position. It will be used in the evaluation of the *covers80* results, as it is the measure most commonly used to compare results for this dataset. 'Recall at 5' is the fraction of queries for which the correct cover is returned in the top 5 ranked results. It is included to give an impression of the performance that could be gained if the current system were complemented with a good alignment-based approach to sort the top-ranking candidates, as proposed by [198], among others. Mean Average Precision (MAP), a more standard evaluation measure, will be used for the *translations* dataset.[7] To compute it, the candidates' ranks r are used to obtain the reciprocal rank r^{-1} for each relevant document returned. Since the datasets in these experiments only contain one relevant document to be retrieved for each query, precision and reciprocal rank are the same, and the mean Average Precision can simply be obtained by taken the mean of r^{-1} over all queries.

In the *translations* dataset, every song that is part of a cover pair is used as a query, and the candidate set always consists of all the other songs. In the *covers80* dataset, a fixed list of 80 queries and 80 candidates is maintained. A random baseline was established for this configuration at MAP = 0.036 for the *translations* dataset, with a standard deviation of 0.010 over 100 randomly generated distance matrices. In *covers80*, less songs are used as a retrieval candidate. Random baselines were found at R_1 = 0.012 (0.013), R_5 = 0.060 (0.026).

7 The difference is due to the a change in small change in implementation between the experiments and the availability of different baselines for each of the datasets.

Experiment 1: Global Fingerprints

The first experiment involves a straightforward evaluation of a few feature combinations using the *covers80* dataset. The three descriptors were extracted for all 160 complete songs. Pitch contours were computed using Melodia and chroma features using HPCP, with default settings [168].[8] For efficiency in computing the pitch bihistogram, the pitch contour was median-filtered and downsampled to $1/4$ of the default frame rate. The bihistogram was also slightly compressed by taking its point-wise square root.

As we observed that key detection was difficult in the present corpus, the simplest key handling strategy was followed: features for a query song in this experiment were not aligned to any tonal center. Instead, each query is transposed to all 12 possible tonics, and the minimum of the 12 distances to each other fingerprint is used to rank candidates.

All representations (*PB*, *CC* and *HA*) were then scaled to the same range by normalizing them for each fragment (subtracting the mean of their n dimensions, and dividing by their standard deviation; $n = 144$). In a last step of the extraction stage, the features were scaled with a set of dedicated weights $w = (w_{PB}, w_{CC}, w_{HA})$ and concatenated to 432-dimensional vectors. We refer to these vectors as the *global fingerprints*.

The main experiment parameters for the features described above are d, the look-ahead window of the bihistogram *PB*, and the weighting w of each of the three features when all are combined. Parameter d was found to be optimal around 0.500 s. Figure 22 shows the pitch bihistogram, chroma correlation, and harmonization descriptors in matrix form for an audio fragment from the *translations* dataset.

Experiment 2: Thumbnail Fingerprints

In a second experiment, the songs in the database were first segmented into structural sections using structure features, as described by Serrà [175]. This segmentation approach performed best at the

8 see section 2.1.4 and `mtg.upf.edu/technologies`

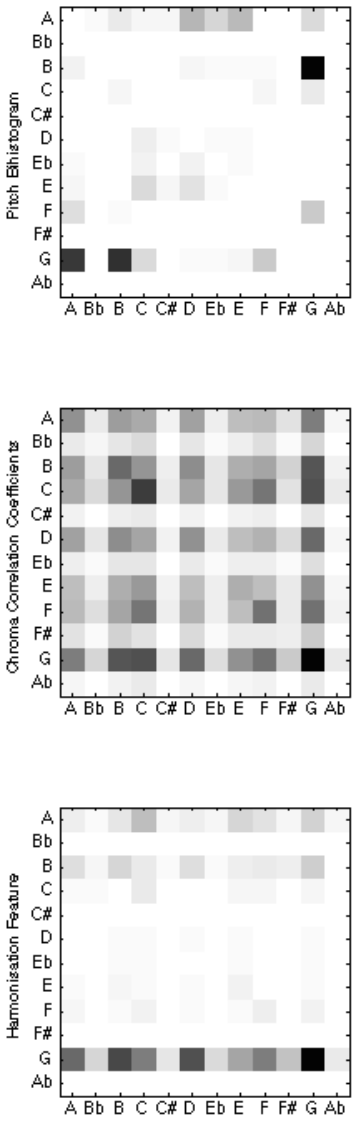


Figure 22.: An example of the pitch bihistogram, chroma correlation, and harmonization descriptors for an audio fragment from the *translations* dataset (in matrix form, higher values are darker). The pitch bihistogram at the top shows how pitch classes A and B appear closely after G, and G after B.

2012 MIREX evaluation exchange in the task of 'music structure segmentation', both for boundary recovery and for frame pair clustering. (A slight simplification was made in the stage where sections are compared: no dynamic time warping was applied in our model, assuming constant tempo.) From this segmentation, two non-overlapping thumbnails are selected as follows:

1. Simplify the sequence of section labels (e.g. ababcabcc): merge groups of section labels that consistently appear together (resulting in AAcAcc for the example above).
2. Compute the total number of seconds covered by each of the labels A, B, C... and find the two section labels covering most of the song.
3. Return the boundaries of the first appearance of the selected labels.

The fingerprint as described above are computed for the full song as well as for the resulting thumbnails, yielding three different fingerprints: one global and two *thumbnail fingerprints*, stored separately. As in experiment 1, we transposed query thumbnails to all keys, resulting in a total of 36 fingerprints extracted per query, and 3 per candidate.

Experiment 3: Stability Model

In the last experiment on the *covers80* data, we want to show that the interpretability of the descriptors allows us to easily incorporate musical knowledge.

In [18], a collaboration with Dimitrios Bountouridis, a model of stability in cover song melodies was introduced. The model was derived independently of these experiments, through analysis of a separate dataset of transcribed melodies of cover songs variations, the *Cover Song Variation* dataset. The dataset transcriptions of 240 performances of 60 distinct song sections from 45 song. It includes four or more performances of each section, as described in [17].

Given the melody contour for a song section, the model estimates the stability for each note in the melody. Stability is defined as the

probability of the same pitch appearing in the same place in a performed variation of that melody. The empirical stability is based on multiple sequence alignment of melodies from the database.

The stability estimates produced by the model are based on three components that are found to correlate with stability: the duration of notes, the position of a note inside a section, and the pitch interval. The details of the model and its implementation are described in [18]. As an example, figures 24 and 25 show how stability relates to pitch interval and position in a section.

From these three findings, two were integrated in the cover detection system. The stability vs. position curve (with position scaled to the $[0, 1]$ range) was used as a weighting to emphasize parts of the melody before computing the thumbnails' pitch bihistogram. The stability per interval (compressed by taking its square root) was used to weigh the pitch bihistogram directly. (Note that each bin in the bihistogram matrix corresponds to one of 12 intervals.) The trend in duration is weak compared with the other effects, so is not used in the experiments in this study.

Experiment 4: The translations *Dataset*

In experiment 4, we apply the global fingerprints method to a real music heritage collection, the *translations* dataset. This experiment is performed to find out what range of accuracies can be obtained when a dataset is used that better represents the musicology scenario described in the beginning of this chapter (5.1.2).

Key handling is approached differently here: fingerprints are transposed to a common tonal center, as found using a simple key-finding algorithm. A global chroma feature is computed from a full chroma representation of the song. This global profile is then correlated with all 12 modulations of the standard diatonic profile to obtain the tonic. The binary form (ones in the 'white key' positions and zeros in the others) is used, as in figure 23.[9]

9 Note that this doesn't assume that a melody is in major: minor key melodies simply get aligned to their third scale degree, as suggested earlier.

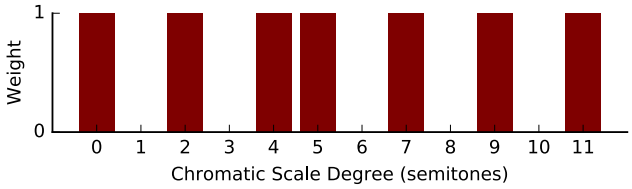


Figure 23.: Diatonic pitch profile used for key handling in the *translations* dataset.

5.3.3 *Results & Discussion*

Table 2 summarizes the results of the all four experiments.

Experiment 1

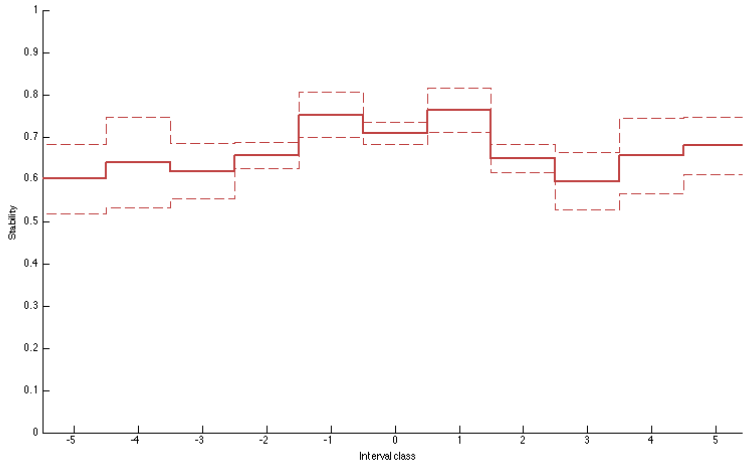
In Experiment 1, each descriptor was first tested individually (only one of w_B, w_C, w_H is non-zero). Results for h, a simple 12-dimensional melodic pitch histogram, and g, a harmonic pitch histogram (chroma summed across time), are added for comparison. They set a strong baseline of around R_1 = 16%–18%. From the newly proposed features, the harmony descriptors (chroma correlation coefficients) perform best, with an accuracy of over 30%, and when looking at the top 5, a recall of 53.8%.

After performing a minimal grid search with weights in $\{1, 2, 3\}$, it is found that, when the three new features are used together, R_1 and R_5 improve slightly. The chroma correlation coefficients contribute most, before the pitch bihistograms and harmonization features.

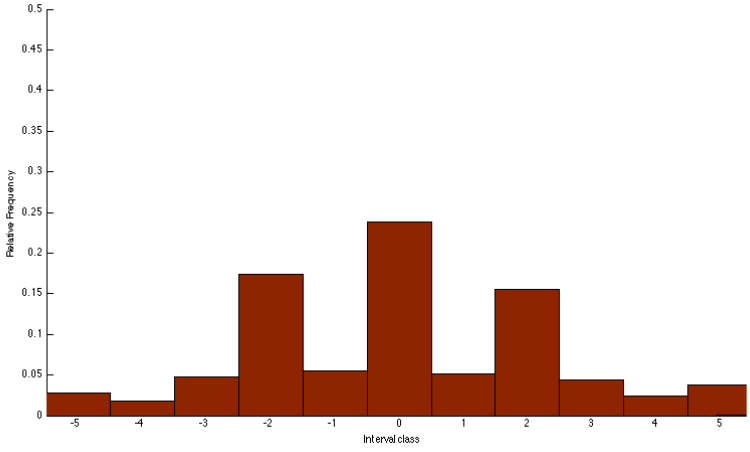
Experiment 2

Results for experiment 2 show that the global fingerprints outperform the thumbnail fingerprints (42.5% vs. 38.8%), and combining both types does not increase performance further.

In less optimal other configurations, it was observed that thumbnail fingerprints sometimes outperformed the global fingerprints, but

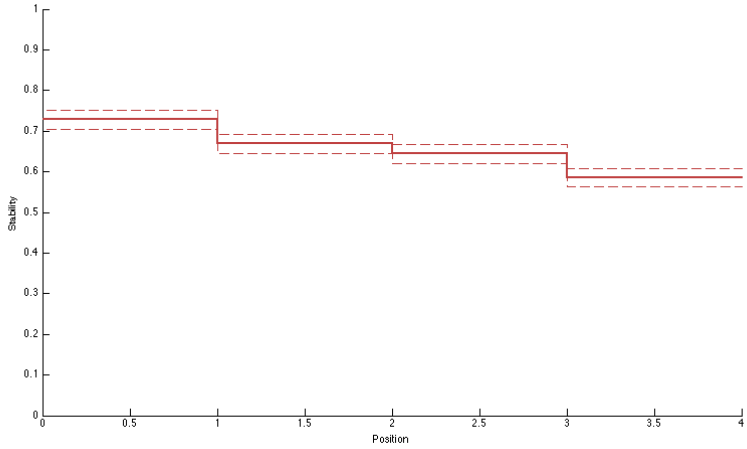


(a) Stability of notes by preceding pitch interval

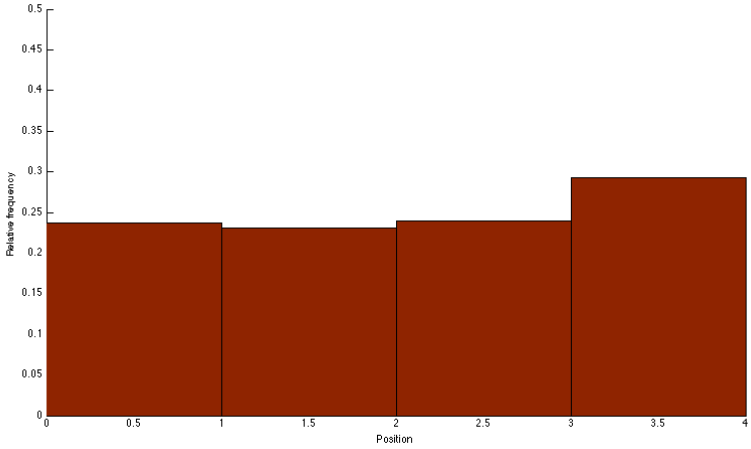


(b) Histogram of pitch interval

Figure 24.: Mean stability of melody notes as a function of pitch interval. Adapted from [18].



(a) Stability of notes by position in the section (scaled to $[0, 1]$)



(b) Histogram of note positions

Figure 25.: Mean stability of melody notes as a function of position in a section. Adapted from [18].

this result didn't generalize to the more optimal configuration with weights $w = (2, 3, 1)$. It is difficult to tell, at this moment, whether the relatively poor performance of the thumbnailing strategy is due to an advantage in capturing all of the songs pitch patterns in one representation, or to poor segmentation.

Experiment 3

When the stability model is integrated in the thumbnail fingerprints, top 1 accuracy reaches 45.0%. This result can be situated between precisions reported using the first alignment-based strategies (Ellis, 42.5%) and a recent scalable system (Walters, 53.8%), see table [198].

This justifies the conclusion that the descriptors proposed in section 5.2.1 capture enough information to discriminate between individual compositions, which we set out to show.

The straightforward embedding of domain knowledge from external analyses further attests to the potential in optimising the proposed representations and fully adapt them to the scalable cover song detection problem.

Experiment 4

Using the *translations* dataset, the precision obtained using just pitch histograms (h) is again added for comparison, and the result is substantial, about 0.27.

However, using just the pitch bihistogram (PB) feature, a MAP of around 0.43 can be obtained, compared to a very competitive 0.42 for using just the chroma correlations (CC). When these features are combined, the MAP goes up to 0.53. In the latter configuration, 44 of the 100 queries retrieve their respective cover version in first place, or in other words, $R_1 = 0.44$, comparable to the accuracy in *covers80*.

The evaluation results for two existing cover detection system, evaluated by Ralf van der Ham for the *translations* dataset, are also included in the table [194]. Van der Ham evaluated Ellis' seminal algorithm based on cross-correlation and Serrà's alignment-based algorithm that is currently considered state-of-the-art [49, 178]. The pro-

posed descriptors are not only faster, but, as can be seen from Table 2, also more powerful than Ellis' cross-correlation method. Serrà's method is much slower, but far superior in performance.

In short, results for the *translations* dataset are not as good as state-of-the-art alignment-based methods, but fairly good for a scalable approach. Specifically, while the dataset poses some extra challenges to Ellis' method, performance for the proposed descriptors is on par with performance in the *covers80* dataset. This justifies the use of the proposed descriptors for the description of older popular music.

5.4 CONCLUSIONS

In this chapter, six new audio descriptors were proposed for the description of harmony and melody. Inspired by notion, from symbolic music analysis, of pitch and interval bigrams, we refer to them as 'audio bigrams', and distinguish (for now) between two kinds: pitch-based audio bigrams and pitch interval-based audio bigrams. Interpretations of the new pitch descriptors were discussed, and their descriptive power is tested in a cover song retrieval experiment.

Performance figures for the experiments, though not state-of-the-art, are a strong indication that the pitch bihistogram feature, the chroma correlation coefficients and the harmonization feature capture enough information to discriminate between individual compositions, proving that they are at the same time meaningful and informative, a scarce property in the MIR feature toolkit.

To illustrate the benefit of the features' simplicity and straightforward interpretation, an independent model of cover song stability has been successfully integrated into the system. Finally, the main findings were confirmed in a cover detection experiment on the *translations* dataset, a dataset of older popular music. In the next chapter, a generalized formulation of the audio bigram paradigm will be proposed.

covers80 dataset experiments	Descriptor	R_1	R_5
Random baseline		0.012	0.060
Ellis, 2006 [48]		0.425	
Ellis, 2007 [48]		0.675	
Walters, 2012 [198]		0.538	
Exp. 1: global fingerprints	h	0.188	0.288
	g	0.163	0.363
	$PB \Leftrightarrow w = (1,0,0)$	0.288	0.438
	$CC \Leftrightarrow w = (0,1,0)$	0.313	0.538
	$HA \Leftrightarrow w = (0,0,1)$	0.200	0.375
	$w = (2,3,1)$	0.425	0.575
Exp. 2: thumbnail fingerprints	$w = (2,3,1)$	0.388	0.513
Exp. 2: global + thumbnail fingerprints	$w = (2,3,1)$	0.425	0.538
Exp. 3: both fingerprints + stability model	$w = (2,3,1)$	0.450	0.563

translations dataset experiments	Descriptor	MAP
Random baseline		0.04
Ellis, 2006 [48]		0.40
Serrà, 2012 [178]		0.86
Exp. 4: global fingerprints only	h	0.27
	$PB \Leftrightarrow w = (1,0,0)$	0.43
	$CC \Leftrightarrow w = (0,1,0)$	0.42
	$HA \Leftrightarrow w = (0,0,1)$	0.30
	$w = (1,1,0)$	0.53

Note. Performance measures are *recall at 1* (R_1; proportion of covers retrieved 'top 1') and *recall at 5* (R_5; proportion of cover retrieved among the top 5) for the *covers80* dataset and MAP for the *translations* dataset. w are feature weights. h = pitch histogram, PB = pitch bihistogram, CC = chroma correlation coefficients, HA = harmonization.

Table 2.: Overview of cover song experiment results.

6

AUDIO BIGRAMS

6.1 INTRODUCTION

In chapter 2 and chapter 5, we argued that alignment algorithms are not a good solution for large-scale cover song retrieval: they are slow, and a database query based on pairwise comparison of songs takes an amount of time proportional the size of the database.

Starting from this observation, and inspired by audio fingerprinting research, we proposed in section 2.2.3 the umbrella MIR task of 'soft audio fingerprinting', which includes any scalable approach to the content-based identification of music documents. Soft audio fingerprinting algorithms convert fragments of musical audio to one or more fixed-size vectors that can be used in distance computation and indexing, not just for traditional audio fingerprinting applications, but also for retrieval of cover songs and other variations from a large collection. We reviewed the most important systems that belong to these categories. Unfortunately, for the task of cover song detection, none of the systems have reached performance numbers close to those of the alignment-based systems.

This chapter provides a new perspective on soft audio fingerprinting. Section 6.2 presents a birds-eye view of research done on the topic. Next, section 6.3 identifies and formalizes an underlying paradigm that allows to see several of the proposed solutions as variations of the same model. From these observations follows a general approach to pitch description, which we will refer to as *audio bigrams*, that has the potential to produce both very complex and very interpretable

musical representations. Finally, section 6.4 present PYTCH, a Python implementation of the model that accommodates several of the reviewed algorithms and allows for a variety of applications, alongside an example experiment that illustrates its use. The toolbox can be used for the comparison of fingerprints, and for song and corpus description in general. It is available online and open to extensions and contributions.

6.2 SOFT AUDIO FINGERPRINTING

This section reviews the various music representations and transformations on which the most important soft audio fingerprinting techniques are based. The systems can be classified into four main groups: spectrogram-based approaches, constant-Q-based approaches, chroma-based approaches, and approaches based on melody.

Spectrogram-based Fingerprinting

Wang and Smith's seminal 'landmark-based' strategy, reviewed in section 2.2.3 is centered on the indexing of pairs of peaks in the spectrogram. In other words: the representation used by the system is the spectrogram, the transformations that are applied are peak detection and pairing of the resulting vectors into a length-4 array.

Constant-Q-based Fingerprinting

At least three systems in the literature have successfully applied the same idea to spectrograms for which the frequency axis is divided into logarithmic rather than linearly spaced bins [51, 181, 192]. Such spectral representations are generally referred to as constant-Q transforms (section 2.1.1) All three systems aim to produce fingerprints that are robust to pitch shifting, an effect that is often applied by DJ's at radio stations and in clubs. The system by Van Balen specifically aims to identify cases of sampling, a compositional practice in which pieces of recorded music are transformed and reused in a new work.

In each case, the idea is that the logarithmic, constant-Q spectrogram preserves relative pitch (also logarithmic in frequency). Transformations used are again peak-detection, and the joining of peaks into pairs [51, 192] and triplets [181].

Chroma-based Fingerprints

The first to perform a truly large-scale cover song retrieval experiment, Bertin-Mahieux and others have taken the landmarks idea and applied it to chroma representations of the audio. Peaks in the beat-aligned chroma features are again paired, into 'jumpcodes'. In an evaluation using the very large *Million Song Dataset* (MSD, see section 3.3.2), the approach was found to be relatively unsuccessful [12, 13].

A follow-up study by Bertin-Mahieux also uses chroma features, but follows an entirely different approach. Beat-aligned chroma features are transformed using the two-dimensional discrete Fourier transform (DFT), to obtain a compact summary of the song that is somewhat hard to interpret, but roughly encodes recurrence of pitch intervals in the time domain [13]. Since only the magnitudes of the Fourier coefficients are used, this '2DFTM' approach is robust to both pitch shifting and tempo changes by design. Results are modest, but formed the state-of-the art in scalable cover song retrieval at the time.

Humphrey et al. have taken this idea further by applying a number of feature learning and dimensionality reduction techniques to the above descriptor, aiming to construct a sparse geometric representation that is more robust against the typical variations found in cover songs [75]. The method performs an initial dimensionality reduction on the 2DFTM features, and then uses the resulting vectors to learn a large dictionary (using k-means clustering), to be used as a basis for a higher-dimensional but sparser representation. Finally, supervised Linear Discriminant Analysis (LDA) is used to learn a reduced embedding optimized for cover song similarity. Their method achieves an increase in precision, but not in recall.

Other soft audio fingerprinting systems are simpler in concept, e.g., the procedure proposed by Kim et al. [90]. This system takes 12-

dimensional chroma and delta chroma features and computes their 12×12 covariance matrices to obtain a global fingerprint for each song. This effectively measures the co-occurrence of pitch energies and pitch onsets, respectively. This relatively simple strategy achieves good results on a small test set of classical music pieces. A later extension by the same authors introduces the more sophisticated 'dynamic chroma feature vectors', which roughly describe pitch intervals [89].

A very similar family of fingerprints was proposed in chapter 5. The chroma correlation coefficients (equation 40) and harmonic interval co-occurrence (equation 48) descriptors follow a covariance-like transformation to compute co-occurrence.

Melody-based Fingerprints

A second set of fingerprint functions proposed in chapter 5 also makes use of melody information, as extracted by a melody estimation algorithm: the pitch bihistgram (equation 39), melodic interval bigrams (equation 46), and the harmonization and harmonization intervals descriptors (equations 43, 49).

The first two aim to measure ordered co-occurrence in melodic pitch—counting co-occurrence of pitch class activations given a certain maximum offset in time. The latter describe the co-occurrence of harmonic and melodic pitch.

6.3 UNIFYING MODEL

6.3.1 *Fingerprints as Audio Bigrams*

The fingerprinting model we propose in this chapter builds on the following observation:

> Many of the fingerprinting methods listed in the previous section detect salient events in a time series, and pair these over one or more time scales to obtain a set or distribution of bigram-like tokens.

The paradigm identified here will be referred to as the *audio bigram* paradigm of fingerprinting. We propose the following definition:

> Audio bigrams are pairs of salient events, extracted from a multidimensional time series, that co-occurr within a specified set of timescales.

Audio bigrams and distributions over audio bigrams can be used as fingerprints.

This definition will be further formalized in the next section. However, it may already become apparent how some of the example systems from section 6.2 can be mapped to the above formulation of the audio bigrams model.

In Wang's landmark approach [199], for example, the salient events are peaks in the linear spectrum (X = DFT magnitudes), and the time scales for pairing are a set of fixed, linearly spaced offsets τ up to a maximum horizon Δt (τ and Δt in frames).

$$\tau = \{1, 2 \ldots \Delta t\} \tag{51}$$

yielding a set of Δt peak bigram distributions for the total of the fragment.

The constant-Q-based variants [51,192] and the jumpcodes approach [12] reviewed above are precisely analogous, with salient events as peaks in the logarithmic spectrum (X = CQT) and beat-aligned chroma features, respectively.

For the case of the melody bigrams, the salient events are pitch class activations pertaining to the melody and the time scale for pairing is a single range of offsets

$$\tau \leq \Delta t. \tag{52}$$

Chroma (and delta chroma) covariance and correlation matrix features are even simpler under this paradigm: pitch activations are only paired if they are simultaneous, i.e. $\tau = 0$.

Two approaches that don't seem to be accommodated at first sight, are Bertin-Mahieux and Humphrey's algorithms based on the 2D DFT. In the remainder of this section, we will show that:

1. a formulation of the audio bigram model exists that has the additional advantage of easily being vectorized for efficient computation,
2. the vectorized model is conceptually similar to the 2D Fourier transform approach to cover song fingerprinting,
3. the model is closely related to convolutional neural network architectures and can be used for feature learning.

It is good to point out that the model will not accommodate all of the algorithms completely. Notably, in the adaptation of landmark-based fingerprinting as described here, some of the time information of the landmarks is lost, namely, the start time of the landmarks. We believe this can ultimately be addressed,[1] but currently don't foresee any such adaptation, as the primary aim at this stage is to explore and evaluate the commonalities between the algorithms.

6.3.2 *Efficient Computation*

In this section, we further formalize the model and characterize its computational properties by proposing an efficient reformulation. The first step to be examined is the detection of salient events.

Salient Event Detection

In its simplest form, we may see this component as a peak detection operation. Peak detection in a 2-dimensional array or matrix is most simply defined as the transformation that marks a pixel or matrix cell as a peak (setting it, say, to 1) if its value is greater than any of the values in its immediate surroundings, and non-peak if it's not (setting it to 0).

Peak detection may be vectorized using dilation. Dilation, denoted with $\oplus$, is an operation from image processing, in which the value of

1 e.g. by not extracting one global fingerprint, but fingerprinting several overlapping segments and pooling the result, cf. [13,75]

a pixel in an image or cell in a matrix is set to the maximum of its surrounding values. Which cells or pixels constitute the surroundings is specified by a small binary 'structuring element' or masking structure.

Given a masking structure S_m, the dilation of the input matrix X with S_m is written as $S_m \oplus X$. The peaks, then, P are those positions in X where $X \geq S_m \oplus X$. In matrix terms:

$$P = h(X - S_m \oplus X) \tag{53}$$

where

$$h(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{54}$$

the Heaviside (step) function.

As often in image processing, convolution, denoted with $\otimes$, can be used as well. We get:

$$P = h(X - S_m \otimes X) \tag{55}$$

where $h(x)$ as above, or if we wish to retain the peak intensities,

$$h(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{56}$$

the rectification function.

Equivalently, we may write

$$P = h(S \otimes X) \tag{57}$$

where S is a mostly negative kernel with center 1 and all other values equal to $-S_m$, similar to kernels used for edge detection in images (top left in figure 26).

The latter approach, based on convolution, allows for the detection of salient events beyond simple peaks in the time series. As in image processing and pattern detection elsewhere, convolutional kernels can be used to detect a vast array of very specific patterns and structures ranging, for this model, from specified intervals (e.g. fifths or sevenths) over major and minor triads to 'delta-chroma' and particular interval jumps. See figure 26 for simplified examples.

$$\begin{bmatrix} -1 & -1 & -1 & -1 & -1 & -1 \\ 17 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 2 & -1 & -1 & -1 \\ 2 & -1 & 2 & -1 & -1 & -1 \\ 2 & -1 & 2 & -1 & -1 & -1 \end{bmatrix}$$

$$\begin{bmatrix} 5 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 5 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Figure 26.: Examples of event-detecting kernels. Rows are time frames, columns can be thought of as pitch classes or frequency bins. They roughly detect, respectively, edges or single peaks, a two-semitone interval sounding together, a two-semitone jump, and 'delta chroma'.

The above is, as is said, a 'vectorized' algorithm, meaning that it is entirely formulated in terms of arrays and matrices. When all computations are performed on arrays and matrices, there is no need for explicit *for* loops. For practical implementations, this means that, in high-level programming languages like Matlab and Python/Numpy, code will run significantly faster, as the *for* loops implied in the matrix operations are executed as part of dedicated, optimized, lower-level code (e.g., C code). Additionally, convolution can be executed very efficiently using FFT.

Co-occurrence Detection

Co-occurrence can be formalized in a similarly vectorized way. Consider that the correlation matrix of a multidimensional feature can be written as a matrix product:

$$F = P^\top \cdot P \tag{58}$$

provided each column in P has been normalized by subtracting of the mean and dividing by the standard deviation. (When only the subtraction is performed, F is the covariance matrix). If P is a chroma feature, for example, the resulting fingerprint measures the co-occurrence of harmonic pitch classes.

When a certain time window for pairing needs to be allowed, things get a little more complicated. We propose an efficient approach in which dilation or convolution is again applied prior to the matrix multiplication. In this case, the structuring element we need is a binary column matrix (size along the pitch dimension is one) of length $2\Delta t + 1$, denoted T.

$$T = [\underbrace{0 \quad \dots \quad 0}_{\Delta t} \quad 0 \quad \underbrace{1 \quad \dots \quad 1}_{\Delta t}]^\top \tag{59}$$

The co-occurrence feature can then be defined as

$$F = P^\top \cdot (T \oplus P). \tag{60}$$

where P may contain, for example, the melody, $M(t, p)$, a chroma-like matrix containing 1 when a pitch class p is present at time t in the melody and 0 everywhere else.

To see how the above F is mathematically equivalent to the proposed co-occurrence matrix, consider

$$P' = (T \oplus P). \tag{61}$$

By definition of $\oplus$,

$$P'(t, i) = \max_\tau (P(t + \tau, i)) \quad \tau = 1 \dots \Delta t \tag{62}$$

so that

$$F(i, j) = \sum_t P(t, i) \ \max_\tau (P(t + \tau, j)) \quad \tau = 1 \dots \Delta t \tag{63}$$

which, for a binary melody matrix $M(t, p)$, translates to

$$B(p_1, p_2) = \sum_t M(t, p_1) \ \max_\tau (M(t + \tau, p_2)) \tag{64}$$

the definition of the pitch bihistogram as given in section 5.2.1 (equation 39).

Assuming that the melody matrix M is based on the melody $m(t)$, this also translates to

$$F(i, j) = \sum_t \max_\tau \begin{cases} 1 & \text{if } m(t) = i \text{ and } m(t + \tau) = j, \\ 0 & \text{otherwise,} \end{cases} \tag{65}$$

the standard definition of the co-occurrence matrix over discrete one-dimensional data.

Alternatively, convolution can be applied, and we get

$$F = P^\top \cdot (T \otimes P) \tag{66}$$

or, in terms of $m(t)$,

$$F(i,j) = \sum_t \sum_\tau \begin{cases} 1 & \text{if } m(t) = i \text{ and } m(t+\tau) = j, \\ 0 & \text{otherwise,} \end{cases} \tag{67}$$

provided S is again binary.

The difference between these two types of co-occurrence matrix is small for sufficiently sparse M, in which case $\max_\tau \approx \sum_\tau$. This is generally true for natural language data, a context in which co-occurrence is often used. It also holds for the sparse peak constellations used in classic landmark-based audio fingerprinting. For more general, dense matrices, the convolution-based F will scale with the density of M while the dilation-based F will not. When efficiency is important, the convolution-based approach should be preferred, so there is an advantage in enforcing sparsity in P.

Methods such as landmark extraction, that do *not* sum together all co-occurrences for all offsets $\tau \leq \Delta t$, can be be implemented using multiple T_k, each of the form:

$$T_k = [\underbrace{0 \quad \dots \quad 0}_{k} \quad 0 \quad \underbrace{0 \quad \dots \quad 0 \quad 1}_{k}]^\top \tag{68}$$

These then yield a set of F_k, one for each offset $k = 1, 2 \dots K$. Naturally, more complex T_k are possible as well.

We conclude that the pairing of salient events over different time scales can be completely vectorized for efficient computation using image processing techniques such as dilation, convolution or both.

Summary

Combining the above gives us a simple two-stage algorithm for the extraction of audio bigram distributions.

Given an input times series X (time frames as rows), a set of n masking structures $\{S_i\}$ and a set of K structural elements $\{T_k\}$ specifying the time scale for co-occurrence, we apply

1. *salient event detection* using
   - convolution with S_i:

$$X'_i = S_i \otimes X \tag{69}$$

   - rectification:

$$P(i,t) = h(X'_i(t)) \tag{70}$$

   $i = 1 \ldots n$.

2. *co-occurrence detection* using
   - convolution with T_k:

$$F_k(i,j) = P^\top \cdot (T_k \otimes P) \tag{71}$$

   $i, j = 1 \ldots n$ and $k = 1 \ldots K$.
   - optional normalization.

so that $F_k(i,j)$ in the matrix form of fingerprint F_k encodes the total amount of co-occurrences of S_i and S_j over the time scale specified by T_k.[2]

6.3.3 *Audio Bigrams and 2DFTM*

We have already shown how pitch bihistogram, chroma covariance and chroma correlation coefficients can be implemented using the above algorithm. Implementing delta chroma covariance, as in [90],

2 The above example assumes convolutions are used, and the further restriction that all S_i have a number of columns equal to that of X, so that each convolution yields a one-dimensional result. Cases where the number of rows in S is smaller are equivalent to using a set S_i of i shifted filters, where i is the difference in number of columns between S and X.

is just as easy: the difference filter shown in figure 26 (bottom right) is applied to the chroma before computing co-occurrence, everything else is the same. Spectrogram, constant-Q and chroma landmarks are also straightforward to implement: S is a peak detection kernel and $T_k = e_k, k = 1 \ldots K$.

Can this formalization also be linked to the 2DFTM and feature learning approaches by Bertin-Mahieux et al. and Humphrey et al.? The most straightforward intuition behind the output of the 2D Fourier magnitude coefficients over a patch of chroma features, is that it encodes periodic patterns of pitch class activation in time.

The model proposed in this chapter measures co-occurrences of events given a set of timescales. In other words, its aspires to do just the same as the 2DFTM-based systems, but with a constraint on what kinds of recurrence in time and pitch space are allowed, by linking it to the bigram paradigm that has been successful in other strands of audio fingerprinting.

Audio Bigrams and Convolutional Neural Networks

The above set of transforms is very similar to the architecture of convolutional neural networks as used in computer vision and artificial intelligence.

Convolutional neural networks (CNN) are a class of artificial neural networks in which a cascade of convolutional filters and non-linear activation functions is applied to an input vector or matrix (e.g. an image). Common non-linear functions include sigmoid functions (e.g. tanh) and the rectification function, used in so-called rectified linear units or ReLU's.

CNN are much like other neural networks, in that most of the layers can be expressed as a linear operation on the previous layer's output followed by a simple non-linear scaling of the result. The coefficients of these linear operations can be seen as the 'weights' of connections between neurons, and make up the majority of the network's parameters. These parameters can typically be learned given a large dataset of examples. An important advantage of CNN over other neural net-

works is that the connections are relatively sparse, and many of the weights are shared between connections, both of which make learning easier. However, learning these parameters in the context of variable-length time series presents an extra challenge: the dimensions of the output and the the number of weights of a typical neural network are assumed to be constant.

The audio bigram model, as it is summarized in section 6.3.2, only consists of convolutions, one non-linear activation function h and a dot product. This makes it a type of convolutional neural network, and a rather simple one: there are relatively few parameters. Conveniently, and to our knowledge, unprecedentedly, the audio bigram approach circumvents the variable-length input issue by exploiting the fixed size of the dot product in Equation 71. The correspondence between audio bigrams and CNN's suggests that, for a given task, optimal matrices S_i and T_k may perhaps be learned from a sufficiently large collection of matches and mismatches.

Audio Bigrams and 2DFTM

Finally, because of the convolution-multiplication duality of the DFT, the audio bigram model can be considered the non-Fourier domain analogue of the k-NN-based system proposed by Humphrey, who describes their system as 'effectively performing convolutional sparse coding' [75].

Future experiments will determine whether standard learning algorithms using back-propagation can be used for this kind of convolutional architecture, and whether an improvement in tasks like sample identification and cover song detection can be achieved using the resulting models.[3]

Figure 27.: Classes (Song, Experiment) and modules (fingerprinting, plotting) in the soft audio fingerprinting toolbox PYTCH.

6.4 IMPLEMENTATION

6.4.1 PYTCH

We provide an implementation of the above ideas in the form of PYTCH, a Python toolbox for soft audio fingerprinting and pitch-based song description available at `www.github.com/jvbalen/pytch`.

The toolbox builds on two primary classes and one important module. A class *Song* contains an ID, a clique ID (denoting the group of covers it belongs to), and any available raw features (e.g., chroma, melody). Centrally in the toolbox is the *fingerprint* module, containing the fingerprinting transforms. It implements a set of fingerprinting methods using some of the feature transforms reviewed out in section 6.2 in the implementation proposed in 6.3. New transforms and new configurations of the existing architecture can be added here. On top of this, a class *Experiment* can be used to evaluate fingerprinting methods, and a *plotting* module can be used for visualization. Figure 27 illustrates the class and module structure of the toolbox.

3 Experiments will be documented at `https://github.com/jvbalen/cover_id`

6.4.2 *Code Example*

In the following (Python) example, an experiment is run on a set of 100 candidates and queries for which the *Song* class has access to a file with chroma base features. This involves three steps.

- Songs and their features are loaded upon construction of the *Collection* objects `queries` and `candidates`.

```
import song

query_ids = range(0,100)
candidate_ids = range(100,200)

file_dir = 'test_corpus\'
file_list = 'list.txt'

queries = [song.Song(file_dir, file_list, id)
           for id in query_ids]
candidates = [song.Song(file_dir, file_list, id)
              for id in candidate_ids]
```

- A fingerprinting function and its parameters are chosen and passed to the object `my_experiment` of the *Experiment* class.

```
import fingerprint as fp
import experiment as xp

fingerprint_type = fp.pitch_bihistogram
fingerprint_params = {'win': 0.5,
                      'normfunction': 'whiten',
                      'keyhandling': 'transpose'}
my_experiment = xp.Experiment(queries, candidates,
                              fingerprint_type,
                              fingerprint_params)
```

- Finally, the experiment is run. The experiment is small enough for the system to compute all pairwise (cosine) distances between the fingerprints.

```
my_experiment.run(dist_metric='cosine',
                  eval_metrics=['map'])
print my_experiment.results
```

In most practical uses of this toolbox, it is advised to set up a new module for each new dataset, overriding the *Song* and *Experiment* constructors to point to the right files given an ID. In the future, an index may be added to store and retrieve fingerprint hashes for large collections.

6.4.3 *Example Experiment*

We now demonstrate in an example experiment how bigram-based fingerprints can be compared, by testing a number of configurations of the system in a cover song retrieval experiment.

As a dataset, we use a subset of the *Second Hand Song* dataset of 1470 cover songs.[4] The subset contains 412 'cover groups' or *cliques*, and for each of these we have arbitrarily selected one song to be the query. The other 1058 songs constitute the candidate collection. Since the subset is based on the second hand song dataset, we have access to pre-extracted chroma features provided by the Echo Nest. Though not ideal, as we don't know exactly how these features were computed (see section 3.4), they make a rather practical test bed for higher-level feature development.

We implemented four bigram-based fingerprints: three kinds of chroma co-occurrence matrices (correlation, covariance, and chroma difference covariance following [90, 191]), and one chroma landmark system, roughly following [12]. The results, with a description of the kernels S and T, are given in Table 3. The chroma landmark strategy was optimized over a small number of parameters: T_k was settled on

4 http://labrosa.ee.columbia.edu/millionsong/secondhand

System	MAP	R_1	R_5
Random baseline	.012	.001	.002
Chroma correlation no S, no T	.181	.097	.155
Chroma covariance no S, no T	.223	.112	.194
Chroma difference covariance $S = [-1, 1]^\top$, no T	.114	.051	.107
Chroma landmarks S simple peak detection T_k of form $[\ldots 0, 1]^\top$	.367	.189	.340

Table 3.: Results table for the example cover song experiment. Chroma landmarks outperform other bigram-type fingerprints.

a set of length-k arrays where $k = 1 \ldots 16$. The best length of the peak-detecting matrix S for the system was found to be 32. Only one peak detection matrix was used.

As can be seen from the table, the chroma landmark system outperforms the other systems. We believe this supports the hypothesis that, starting from the kernels S and T that describe this transform, a more powerful representation can be learned.

6.5 CONCLUSIONS AND FUTURE WORK

We have reviewed a selection of soft audio fingerprinting methods, and described a fingerprinting model that allows to see these methods as variations of the *audio bigram* paradigm. The audio bigram model measures co-occurrence of pre-specified salient events in a multidimensional time series. We have presented an exploration of the computational architecture of the model and showed that it can be implemented as a particular type of convolutional neural network. The

model can therefore be optimised for specific retrieval tasks using supervised learning. Finally, we have introduced an implementation of the model, PYTCH.

As future work, we plan a more extensive evaluation of some of the existing algorithms the system is capable of approximating. Standard datasets like the *covers80* dataset can be used to compare results to existing benchmarks. If the results are close to what the original authors have found, PYTCH may be used to do a comparative evaluation that may include some variants of the model that have not previously been proposed.

We also intend to study the extent to which the convolutional network implementation of the model can be trained, and what kind of variants of the models this would produce. This can be done most easily using the complete *Second Hand Song* dataset, because a rather large number of train and test data will be required.

Part III

CORPUS ANALYSIS OF HOOKS

7

HOOKED

The aim of the next two chapters is to analyze the phenomenon of *hooks*. This chapter presents Hooked, the music game we developed to collect data on the music memory of a large number of frequent popular music listeners in the Netherlands and the United Kingdom. The analysis of the data will be presented in chapter 8.

The first two sections introduce the question that will be addressed, and the necessary backgrounds offered by musicology and music cognition. Sections 7.3 and 7.4 describe the design and implementation of the Internet experiments we set up to collect a dataset of hooks.[1]

7.1 CATCHINESS, EARWORMS AND HOOKS

Music cognition, as introduced in chapter 1, is the science of music listening. It seeks to explain the wide range of processes that are involved when a person listens to music, including memory, attention, expectation and emotion [152].

In the scientific study of popular music, this listener-centered perspective on music and the experiences around it is crucial. Popular music has often been characterized by the length of songs: three to four minutes in most cases. Listening to popular music requires a different kind of attention than, say, symphonic music. Memory is very important too: a popular or particularly engaging pop song is often said to be 'catchy' or 'an earworm', suggesting it has somehow been

1 The names *Hooked!* and *Hooked on Music* will be used to distinguish the implementations from the experiment in its most general form.

etched into memory: popular music is, on a fundamental level, an interaction with music memory.[2]

The notion of catchiness is a fuzzy one and difficult to define, but it is generally understood as 'easily recalled to memory', or 'memorable'. Many listeners confidently discriminate between music they consider catchy, and non-catchy music, even upon first or second listen. In the musicology and music cognition literature, however, the notion of catchiness is rarely analyzed, despite its vital role in music memory and popular music.

There is some literature on the related notions of *hooks* and *earworms*. *Earworms* are songs or parts of songs that get stuck in one's head. Over the last decade, earworms have become a serious subject of research in music psychology and neuroscience, where they are often referred to as involuntary musical imagery, and so have a range of related phenomena, such as musical hallucinations and other, non-musical imagery. Little is known about the kind of songs that easily get stuck in one's head, but we are learning fast about the context in which it happens, and the individual differences that correlate with it [84, 133].

A *hook*, in the songwriting literature and in musicology, is the part of a song that make it catchy. It is the part that grabs the attention; it grabs or 'hooks' the listener, trapping the song into the listeners' memory as the result of its memorability [28]. As with earworms, very little is known about what contributes to the 'memorability' or long term memory salience of popular music, about what makes a hook a hook [187]. In the next section, we define the notion of catchiness and hooks, and review what is known about the musical variables that are associated with the phenomenon.

2 Of course, popular music, on an equally fundamental level, must also be understood in terms of processes of embodied cognition, identity and social dynamics, semiotics... We consider the music's interaction with memory an important facet of the popular music experience, but, recalling Huron's notion on explanatory 'closure', we do not choose the angle of music and memory with the intent of replacing or explaining away other perspectives on popular music—see chapter 1, and [79].

7.2 HOOKS IN MUSICOLOGY AND MUSIC COGNITION

From a cognitive point of view, we can define a hook to be the most salient, easiest-to-recall fragment of a piece of music. Likewise, we can define catchiness as long-term musical salience, the degree to which a musical fragment remains memorable after a period of time. By this definition, every piece of music will have a hook—the catchiest part of the piece—even if some pieces of music have much catchier hooks than others. In principle, a piece of music may also have multiple hooks: two or more fragments of equivalent salience that are nonetheless more salient than all others in the piece.

This definition—the hook as the location of maximum catchiness—makes the notion of hooks useful for applications, too. A retrieval system that knows the location of the most memorable or recurring part of each song has a strong advantage in human-centered context of information retrieval, e.g. when the user of a music search system expects the system to return something they know [69]. The system could also benefit computationally. In a similarity search, it might ignore redundant or less salient song sections, thus limiting the space over which it needs to search. In other words, understanding hooks can be useful not only to music cognition, but also to music information retrieval.

We review the available literature on hooks and catchiness in the remainder of this section, beginning with musicological accounts in section 7.2.1, followed by an overview of relevant music cognition perspectives in 7.2.2.

7.2.1 *Hooks in Musicology*

Burns' Typology of Hooks

The most important and the only article-length musicological study of hooks is a 1987 typology of popular music hooks by Gary Burns. [28]. The paper consists of an exhaustive overview of hook categories, extensively illustrated with examples.

Textual elements	Non-textual elements
musical elements: – rhythm – melody – harmony lyrics	performance elements: – instrumentation – tempo – dynamics – improvisation production elements: – editing – mix – channel balance – signal distortions – effects

Table 4.: Textual and non-textual elements of a music recording that can make up a hook. Adapted from [28].

Adopting a definition by Monaco and Riordan, Burns starts off by defining a hooks as 'a musical or lyrical phrase that stands out and is easily remembered' [131]. He then proposes that popular music hooks lie in the interaction between *repetition*, *variation* (or *modulation*) and *change* in the elements that make up popular music.

Repetition and change are contrasted to conclude that a discourse entirely based on repetition or continuous change cannot be successful while variation or modulation, terms used interchangeably, are put forward as the critical mechanism by which catchy music is composed.

Burns divides the musical elements that constitute a recording into *textual* elements (melody, harmony, rhythm and lyrics) and *non-textual* elements. The textual elements can be seen as that which can be written, as opposed to non-textual elements which are a product of the performance or production. The set of textual and non-textual elements that could make up a hook is then narrowed down to the list shown in Table 4. The remainder of the paper provides examples.

Not all of the examples are cases of modulation and repetition. Other terms also recur frequently, primarily *distinctiveness*, *surprise* and *intertextuality*. The last term is used to denote hooks that are the result of confusing or associating a piece with a specific other work.

Other work

Other work that theorizes the concept of hooks includes a more recent analysis of hooks by Kronengold [99]. Kronengold suggests that the characteristics of hooks vary across genres, and explores the idea that assortments of hook characteristics might constitute a useful definition of genre.

The analysis further centers on the the possible role of 'musical accidents' (mistakes in performance or production) as hooks. To perceive something as an accident requires a high-level understanding of the music, including the context and the intentions of the performer. Such a strong dependency on context and interpretation places this type of hook alongside the narratives in lyrics and personal associations of the listener, aspects of musical cognition that fall outside the scope of this project.

7.2.2 *Hooks and Music Cognition*

Clearly, Burns' 1987 review is written from a popular music research perspective, and only a few times do notions related to cognition show up in the discussion. What are the implications of Burns' systematization when the discussion is read in light of theories in the music cognition domain? The important keywords in this analysis that are shared with the vocabulary of music cognition are *repetition*, *variation-modulation* and *distinctiveness*. We review the prevailing perspective on these topics.

Repetition and Variation

Elizabeth Margulis has written extensively about the role of *repetition* in Western and non-Western music—identifying repetition as a distinctive property of earworms, for example, and studying the effects on human repetition detection of multiple exposures [117, 118]. Results on the latter suggested that not all repetition is the same: within-phrase repetition is more noticeable than separated, between-phrase

repetition . Repetition of units that form a complete segment at some level also seems to be more noticeable than repetition of units that do not.

Results from another study by Margulis suggested that listeners aesthetically prefer slight variations over verbatim repetitions, in two experiments on contemporary Western Art music and eighteenth century Rondos. This supports Burns' initial argument about the appeal of variation as compared to exact repetition [119].

Studying repetition in lyrics, Nunes et al. showed that songs with repeated lyrics have a higher chance of debuting in the Top 40 charts, and an higher chance of reaching the number one position [140]. The correlation is attributed to an increase of processing fluency of the song's lyrics.

Effects of repetition on musical preference also seem consistent with the *mere exposure effect*: listeners tend to show an increased liking of a piece of music as they become more familiar with it. Huron ascribes this effect to our brains rewarding correct predictions: "listeners prefer familiar stimuli not because they are familiar, but because they are predictable" [82]. In other words: we prefer listening to music with repetitions because the repetitions make it predictable. A number of such associations between familiarity and reward have been confirmed in brain imaging studies, e.g., by Pereira et al. [157]. They show that the brain's 'reward circuitry' is more active for song excerpts that were familiar to the participant prior to the experiment. However, we should be careful to generalize from these experiment, as the kind of familiarity studied tend to vary; in [157], it is the result of repeated exposure on the time scales of weeks, months or years, so it is far removed from the familiarity induced by repetitions within a piece of music. The study also shows an increase in reward, not preference.[3]

3 In the part of the experiment based on familiarity ratings, preference ratings was even controlled for, as any relation between preference and familiarity, in a correlational study, is seen as confounding—we tend to be more familiar with the music we like more.

Chart position and preference ratings are not necessarily good measures of catchiness. They are strongly affected by exposure and the many economic, social, demographic and other factors that in turn affect exposure. In other words, the above arguments are worth some consideration, but while the evidence for an effect of familiarity on preference is strong, the evidence for repetition as a driver of preference is incomplete.

Distinctiveness

A study that does aim to measure memorability, is Müllensiefen and Halpern's experiment on the recognition of popular music melodies [134]. Presenting participants with a set of unknown popular song melodies, some once and some more than once, they ask two questions: have you heard this melody before (within the experiment)? And, how would you rate this melody in terms of pleasantness? The two kinds of ratings are used as an explicit and implicit measure of recall, respectively. Evidence from experiments on the mere exposure effect is cited to support the argument that pleasantness ratings can be a more reliable proxy of implicit familiarity than familiarity as measured by 'explicit' ratings. They then investigate the features that predict recognition of the melodies using a 'discovery-driven' statistical analysis. Given a set of features computed from the (symbolic) melodies, a regression model is used to find the latent factors in a set of symbolic features that best predict the two memory scores.

Results show that explicit memory scores ('I have heard this before') are higher for less typical and less complex melodies, as measured by the mean document frequency and productivity of n-grams, respectively (variables from the FANTASTIC toolbox—see [132] for definitions). Implicit memory shares one of these two factors: scores are higher for melodies composed of less typical motives. In terms of complexity, however, the trend is inverted: more complex memories get higher ratings. These findings—for either rating, less typical melodies get higher scores—suggest a link with *distinctiveness*.

In a more general perspective on distinctiveness, Huron presents in [79,80] an overview of the literature on schema selection and expectation in music. Huron highlights three types of expectation:

Schematic expectations
are expectations that arise from the cumulative exposure to music throughout all of one's life. Several schemas may exist in parallel. Upon listening, schemas are selected based on the type of music (eg. its genre) and the context. Schematic expectations have a quick effect.

Veridical expectations
are the expectations that are associated with knowing a particular work. Veridical expectations are slowly triggered as they require a confident recall of the piece.

Adaptive expectations
are dynamically updated and accumulate during listening of the piece itself. They are especially prominent when listening to a work for the first time. First proposed by Meyer [128].

In the context of Müllensiefen's findings in [134], schematic expectations would reflect 'typicality', what is 'common' in popular music. A violation of these expectations occurs whenever atypical, distinctive patterns are encountered. In other words, Müllensiefen provides some evidence for a relationship between distinctiveness and memory salience, perhaps including long term memory salience and hooks. In light of Huron's technical perspective on expectation, this can be read as evidence for an effect of violations of schematic expectation.

These results are valuable clues in the understanding of memorability and catchiness, however, the scope and context of these findings is important, too: they are based on data from a a small group of 34 participants listening to short, synthesized, monophonic melodies. And most importantly, the test only probes recent memory—somewhere on the order of the length of the experiment. The particularities of this kind of musical memory may be quite different from the long-term memory salience of hooks.

7.2.3 *Summary: Hook Types*

The above imperfect, but valuable findings in music cognition point to, on one hand, the importance of repetition, reinforcing adaptive expectations and contributing to greater processing fluency, and on the other hand, distinctiveness (i.e., violations of schematic expectations).

Emphasizing modulation/variation and distinctiveness, Burns' typology allows for both kinds of hooks. As an example of the first, 'vamp' hooks get their catchiness from the continued repetition of a (often very stereotypical) pattern, conforming to schematic expectations, as in e.g., Louie Louie by The Kingsmen (1963), a song built on a short repeated I-IV-V-IV progression. Similarly, many of Burns' examples of instrumentation and production-related hooks are essentially cases of atypical or distinctive sound or timbre, e.g., the use of a sitar in The Beatles' Norwegian Wood (1965).

Alongside these two types, Burns' typology also allows for the combination of both effects: pattern that are repeated, and thereby very representative for the song, but not for the larger body of music the song belongs to, a notion of hooks that may be referred to as an 'identifying motive'. Examples include the repeated I-bII harmonic progression in Jefferson Airplane's White Rabbit (1967).

One recurring and potentially interesting keyword out of Burns' analysis that hasn't been given much attention in the music cognition literature is *surprise*. Surprise is probably most easily understood in terms of Hurons ideas as an effect of adaptive expectations. Adaptive expectations represent the patterns that are established as a song comes along. In this perspective, adaptive expectations may be reinforced by repetition and violated at moments of surprise. A hook that is cited as an example of both distinctive (unusual in absolute terms) and surprising (in relation to the rest of the melody) is Minnie Ripperton's high-pitched vocal effort in Lovin' You (1975).

These many possible types of hooks—repetition, distinctiveness, identifying motive, surprise—directly and indirectly hypothesized in the literature, will be considered in the next chapter. They will not be regarded as strict hypotheses, to be tested individually, but they

are the motivation for an analysis in which approximate expectations are summarized in 'corpus-based features' (section 8.1).

7.3 EXPERIMENT DESIGN

In the work described in this section, we set out to collect a dataset of hooks: properly sourced annotations of the catchiest parts of a set of songs, where catchiness is measured in a way that respects its above definition as long-term musical salience.

7.3.1 *Measuring Recognisability*

Collecting a dataset of hooks presents a number of challenges. First, an appropriate measure of a hook's catchiness must be found. Having defined the hook as the most memorable, easiest-to-recall fragment, we can look at the psychology literature for measures of recall.

In memory retrieval experiments in the domain of psychophysics, *two-alternative forced choice tasks* are a common experiment paradigm. In this set-up, a participant is asked to answer a binary-choice question, and the response is timed. When applied in memory research, a range of memory retrieval models is available that allow for the choice and response time of items and participants to be combined into a single measure of ease-of-recall, e.g. Ratcliff's drift-diffusion model—more on this in chapter 8 [164].

Following this paradigm, we choose as an empirical measure of catchiness the *stimulus drift rate*, i.e., the ease-of-recall as measured using the drift diffusion model of memory retrieval based on response times. Admittedly, this does not cover all of the possible associations that the informal and overloaded term 'catchiness' might have (e.g., any association with the notion of earworms), but it gives us a rigorous and practical means of measuring the phenomenon.

7.3.2 *Games and Music Research*

A second challenge is that the experiment necessarily focuses on long-term memory—anywhere between days and decades, but definitely longer than the hours or even minutes available in a traditional laboratory experiment. We must therefore rely on a different kind of experiment, and on well-known music, memorized before the experiment itself. This is not a challenge in itself—popular music is easy to find. However, even a collection of very well-known popular songs will still contain a lot of music that is unknown to the average participant. To ensure that we can collect enough annotations of each of the items in the collection, a large number of participants is required. The individual listening history of a large group of participants, in turn, may vary widely, so a sufficiently large collection of songs is needed too.

We argue that these requirements—a large number of participants and a large number of songs—make our data collection problem a good candidate for an Internet-based experiment rather than a traditional laboratory experiment.

Serious Games and Games With a Purpose

Research on computational modeling in music tends to rely on music experts to annotate the ground truth data required for model training and evaluation. For some modeling projects, however, experts are unnecessary, or even undesirable (e.g., for music annotations related to mood and emotion or measuring music similarity: annotations for these tasks must reflect the variation in listeners' perception [3]). They require a large amount of annotations from a diverse group of annotators. This makes them good candidates for crowdsourcing. Crowdsourcing—the outsourcing of an automatable task to a virtual crowd of volunteers or paid freelancers—has been a popular and successful strategy for large-scale data collection. Another recent approach to this—a specific kind of crowdsourcing—involves serious games. Serious games are games of which the primary purpose is not

to entertain. They have found applications in healthcare, education, and professional training [3].

Serious games used for data collection are also called 'games with a purpose' (GWAP). GWAP and Internet-based experiments are increasingly seen as a serious alternative to lab-based experiments. They can potentially reach a much larger, more varied and intrinsically motivated participant pool, which contributes to the ecological validity of the collected data [69]. GWAP have already proven successful for certain tasks in MIR and machine learning.

For the above reasons, we have decided to frame the experiment as a game. Given that most listeners enjoy catchy music, research on catchiness and hooks seems naturally suited for the GWAP paradigm. With an appropriate choice of music collection we believe the GWAP format allows us to collect enough data for a data-intensive analysis of hooks, and possible future applications in content-based MIR—e.g., hook retrieval or predicting catchiness.

7.3.3 *Gameplay*

The general experiment design followed in Hooked is, therefore, a timed recognition task in the form of a game, in which, for a large number of songs, different sections are presented to a large number of participants.

The hook data collection experiment is a game with four screens as part of the main game loop.

Recognition Screen

When the game starts, a song fragment plays, and the user is asked: "Do you know this song?". This is the *recognition screen*. Participants have a limited amount of time $r_{\max}$ to answer the question; two buttons, YES or NO are shown as in figure 28. To incentivize speed, participants can earn points proportional to the time they have left. The response time is stored together with a song fragment identifier, this is the *recognition time* r.

Figure 28.: Screenshot of the recognition, mute and verification screens in *Hooked!*

Mute and Verification Screens

In a laboratory context, one could expect the participant to answer the above question honestly. In a game context, however, there is an incentive to answer positively regardless of whether the participant knows the song or not. This is why answers are verified in the next two screens.

After a participant hits YES in the recognition screen, the sound of the song fragment will mute for m seconds, and the player is asked to follow along in their head. The *mute screen* is shown, on which the participant can see how much longer the sound will be muted.

After mute time, the sound comes back up, and the song fragment continues playing. However, the fragment might start again from $r + m$ seconds (as if the sound was genuinely muted), or it might start somewhere else inside the song, at time $r + m + \Delta$ (as if the record was scratched)–with equal probabilities. The offset Δ will be referred to as the *distraction* offset. The *verification screen* is then shown, in which the participant is asked whether the music continued in the right place. Again, the response time v (for *verification time*) is recorded.

In other music trivia games, such as *Song Pop*,[4] players are asked to select the correct title or artist of the song to prove they know it. In Hooked, this approach is intentionally avoided: listeners may know a piece of music rather well without knowing its exact title or the name of the performing artist. Moreover, even for those who do know such trivia, the extra cognitive load in recalling it to memory would have an unpredictable and currently irrelevant effect on response time. Instead, the above verification tasks was inspired by the idea that music listeners who know a song well are able to follow along in their heads for some time after the fragment is muted.

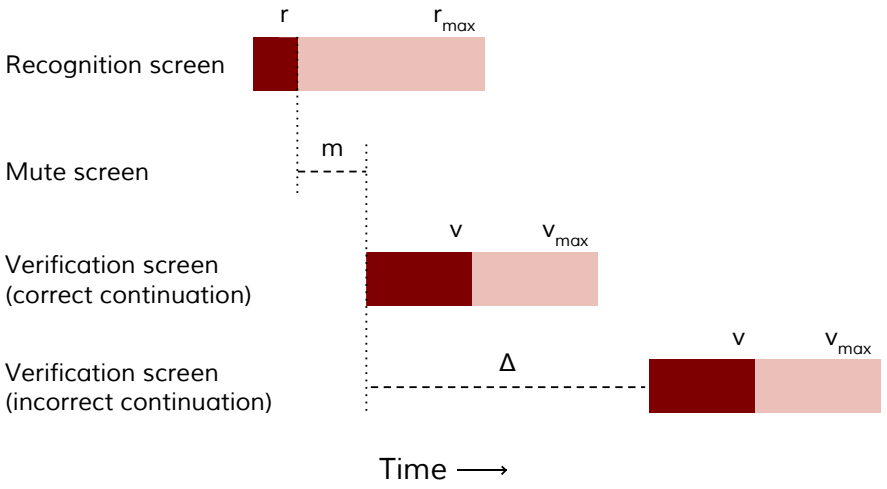
Prediction Screen

Apart from measuring the recognition time as an indicator of ease of recall, we are also interested in participants' intuition on which parts of songs are most catchy. Several times throughout a round of several recognition-verification tasks, Hooked also shows the *prediction screen*. In this screen, participants are asked to listen to two fragments from the same song and choose which they consider catchier—the question reads: 'which fragment is more recognizable'? To give players an extra incentive to give non-random answers, we integrated this task into the game. Each time players complete a prediction task, the chosen fragment is saved. One of the following recognition tasks will then enter a bonus round for double points, presenting the player with the saved fragment that was chosen during the prediction task.

7.3.4 *Experiment Parameters*

The above experiment design involves a number of variables that are recorded and saved to a database as research data: r, v and the answer to all three of the questions (recognition, verification and prediction). Additionally, there are a number of important experiment parameters, including m, $r_{\max}$ and Δ, but also the possible start times of the fragments of the song, which need to be set in advance. Figure 29 shows

4 `http://www.songpop2.com`



Note. The horizontal axis represents time in the song. Vertically separated are three of the four main screens in the game: recognition, mute and verification. In the verification screen, two scenarios are possible: the song continues in the right place, or it continues elsewhere. Audio is playing in the dark red regions.

Figure 29.: Schematic of the parameters and variables in the Hooked experiment.

all variables and parameters in a summary of the time line of the recognition, mute and verification screens of the game.

Start Times

How should start times be chosen? Of course, they could be chosen randomly (e.g., by picking 5 or 7 points in each of the songs in the music collection). However, one of the few agreements in the literature on hooks is that they start at points of considerable structural change [28, 126]. In MIR terms, these would correspond to the boundaries identified by a segmentation system (section 2.2.2). Therefore, start times in Hooked are based on segmentation of the songs by the Echo Nest's structural segmentation algorithm. This not only ensures that

starting points are biased to somewhat meaningful locations inside the song, it also ensures that the resulting data can be aligned with other manually and automatically extracted information for the music in The Echo Nest's collections, as they make up an important resource in MIR.

Maximum Recognition Time, Mute Time and Distraction Offset

In the same literature, there is considerably more debate about the *duration* of hooks, something that must be considered when deciding on the right maximum recognition time $r_{\max}$. While songwriters will often speak of the hook as the entire chorus, only a few seconds are necessary for most listeners to recall a catchy song to memory. One study has shown that after only 400 ms, listeners can identify familiar music with a significantly greater frequency than one would expect from chance [101]. The reality very likely sits somewhere in between, but that still leaves a lot of options.

Another challenge is to find the optimal mute time. The time during which participants follow along in their heads shouldn't be so short that one can judge the continuation on the basis of common-sense musical knowledge or timbral characteristics, leading to a type II error. However, it should also not be too long: results from music memory studies have shown that absolute tempo is part of musical memory, though with errors biased towards higher tempi [108]. This would lead to a type I error.

A third parameter that is difficult to establish theoretically is the distraction offset: different offsets might have a different effect on the difficulty of the task and the resulting type I and type II error rate.

To set these three parameters, a pilot experiment was set up, in which 2 options were tested for each parameter. The maximum recognition time $r_{\max}$ was either 10 s or 15 s, the mute time m was either 2 or 4 s, and the distraction offset Δ in the verification task was either 15 s or -15 s. The pilot consisted of a fully functional version of the game, but with just 160 song sections from 32 songs. Twenty-six participants were recruited.

By asking trusted pilot participants to play part of the game honestly, and part of the game competitively, we were able to deduce an estimate of the expected frequency of type I and type II errors. In a statistical model, these were then compared to the different parameter settings. The results, documented in [23], show a trade-off between type I and type II errors, and surprisingly, evidence that honest playing was rewarded with more points, regardless of the parameter settings. Parameter settings were therefore chosen to maximize the pleasantness of playing honestly: $r_{\max} = 15$ s and $\Delta = +15$ s. Mute time did not have a significant effect, so it was left open as subsequent versions of the game were developed.

7.4 IMPLEMENTATIONS

In this section, two implementations of Hooked will be discussed: *Hooked!* and *Hooked on Music*.

7.4.1 Hooked!

The first version of the Hooked experiment was developed for launch in The Netherlands in December 2013. Named *Hooked!*, it was developed for the iOS mobile operating system (iPhone, iPod Touch and iPad). This decision was made after an investigation into the available options for the on line hosting of the copyrighted music fragments. It is difficult to get permission to let users of a game have access to a collection of music for free. Even with these permissions in place, it is difficult to protect on line audio from being used outside the environment of our game. We were able to circumvent these difficulties by working with Spotify's streaming API. Spotify's paid Premium service allows users to stream music to their mobile devices. The streaming API allows developers to use this service inside their own applications, for users who subscribe to the service. Since integration of this streaming API into iOS applications was much more straightforward than integration under other operating systems, *Hooked!* was only im-

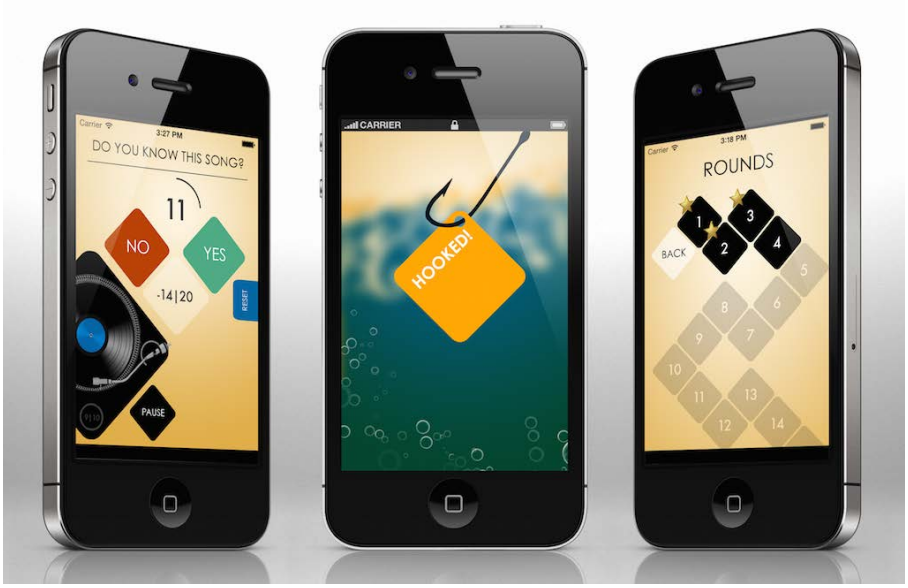


Figure 30.: Visual appearance of the *Hooked!* game. Showing the recognition screen and two more navigation screens (welcome and round selection).

plemented for iOS—realizing that the combined requirements of iOS and Spotify Premium would be a bottleneck for many.[5]

Design

A number of information design aspects of the *Hooked!* game design were completed with advice from Frontwise, an information design agency based in Utrecht, The Netherlands.[6]

A mock-up of the final appearance of the game is shown in figure 30. The recognition, mute and verification screen are shown in more detail in figure 28. In general, the goal in the design process was

5 At the time, however, the market share of iOS was still a lot closer to 50%. Spotify Premium coverage was much lower, but heavily biased to our audience: avid music fans.

6 `https://www.frontwise.com/`

to make the use of *Hooked!* as easy and intuitive as possible. Some of the details that were considered were the color and order of the buttons. E.g.: YES, in the recognition screen, makes the game proceed to a next screen, while NO makes it go back to the recognition screen (but with a different song). Exploiting implicit associations between right-forward and left-backward, the YES button was placed on the right. Were the presentation of instructions (at the beginning of the game) and help text (optionally shown along the way) were also given extensive consideration.

Data

To ensure that the songs used in the game were well-known, the music collection was settled on a subset of the 2012 edition of 'Top 2000', a list of the 'greatest songs of all time' as voted by the listeners of a popular annual radio programme in The Netherlands.[7] As the show is one of the Netherlands' most popular music events of the year, the list represents the musical preferences of a substantial part of the country, including both older and younger generations. The Top 2000 has been characterized as an informal canon of popular music for The Netherlands, though documenting a consensus on popularity rather than quality as judged by experts or the establishment [1].

The advantage of this choice of data is that—unlike with a sample of popular music based on historic chart success—we have some guarantee that the music is still known by many today. On the other hand, any popularity measure based on lists like the Top 2000 has serious drawbacks, too, including somewhat of a demographic and geographic bias, resulting in the absence of some more recent popular styles (e.g., hip hop and electronic dance music, otherwise two very popular genres in the Netherlands). Note, also, that the Dutch audience's preferences in popular music differ from those of other markets (not just because of the presence of music by local artists, but also in the appreciation of particular genres, like hip hop or country

7 `http://www.radio2.nl/top2000`

music). *Hooked!* is therefore suitable only to support models of 'Dutch hooks'—hooks as heard by Dutch popular music listeners.

The *Hooked!* subset of the Top 2000 list included all songs that could be streamed on Spotify, 1591 in total. These songs were divided into 20 groups based on popularity, as measured by The Echo Nest's 'hotttnesss' metric.[8] Each of these groups is used as a level in the game, such that users who finish the game begin (level 1) with the most popular songs and end (level 20) with the lesser-known ones.[9] The resulting unequal distribution of annotations over songs allowed us to collect a sizable number of annotations per song right from the beginning, with more groups of songs collecting annotations as more users joined the game.

Results

After two years, *Hooked!* has been played by 1986 unique players, gathering a total of 167,704 responses to the recognition question. On average, players answered the recognition question positively in 61.5% of the trials, 38.5% of trials were answered negatively or skipped. After a positive answer, the verification question was answered correctly 74% of the time (45.5% of total trials). The top annotated song—Adele's Someone Like You (2011)—was correctly identified 483 times.

7.4.2 Hooked on Music

Building on the experiences developing and publishing *Hooked!*, the COGITCH project partnered with the Manchester Science Festival to produce a more widely accessible version of Hooked, to be launched in 2014. The result, *Hooked on Music*, was a complete reimplementation of the game. It was built primarily in HTML5, with a responsive design that renders on any browser, desktop or mobile, and adapts

8 `http://developer.echonest.com/docs/v4/song.html`

9 It should be noted that the use of the (US-based) Echo Nest popularity data creates a concentration of English language songs at the beginning and Dutch language songs at the end. As a result, many Dutch language songs did not collect enough annotations for further research.

to the size of the screen. Rather than relying on Spotify's streaming API, a music license was negotiated with Dutch collecting societies Buma/Stemra[10] (representing composers and music publishers) and SENA[11] (representing musicians and producers) so that anyone could access the game. Audio was hosted by Soundcloud.[12]

Unlike in *Hooked!*, there are no levels in *Hooked on Music*: for as long as they keep playing, participants are given fragments to recognize, in random order. It also drops the prediction screen in which participants would be asked to predict which of two fragments of a song is more recognizable. The implementation did include a questionnaire through which participants could give us more information about their background (incl. age, level of musical education, estimated weekly hours of music listening). There is also one difference in experiment settings: the mute time (section 7.3.3) in *Hooked on Music* is $m = 4$ seconds. The music used in *Hooked on Music* is a sample of British popular music: 1000 45-second excerpts from 200 chart-topping popular songs from the 1940s to the present (5 excerpts per song).

Over 160,000 participants from 200 countries played *Hooked on Music*, spanning an age range from 15 to 85. A total of over 3 million responses to the recognition question were collected, an order of magnitude more than in *Hooked!*.

7.5 CONCLUSION

In this chapter, we introduced the Hooked experiment on popular music and memory. We presented our motivations to create a dataset of hooks, and reviewed current musicological and cognitive perspectives on the related topics of hooks, distinctiveness and repetition. We then discussed experiment design, explaining our decision to cast the experiment as a game, and two implementations: *Hooked!* and *Hooked on Music*.

10 http://www.bumastemra.nl/en/

11 http://www.sena.nl/en/

12 http://soundcloud.com/

The analysis of the participant data (response times and accuracies) and the audio used in *Hooked!* will be presented in the next chapter. The first iteration of data collection in the *Hooked on Music* game, described above, has only just been closed. The data and audio for this version of the experiment will be analyzed in the months to come (see section 9.3 on current and future work).

8

HOOK ANALYSIS

In this chapter, we build on insights on corpus analysis from chapters 3 and 4, and on the audio bigram features proposed in section 5, to present a corpus analysis of the song recognition data described in chapter 7.

As part of the analysis, we propose a new set of *second-order* features (section 8.1. The notion of second-order music descriptors is inspired by latent semantic analysis methods from text retrieval. They encode typicality and distinctiveness of feature values. By adapting the concept to audio features, we are able to present a cognitively adequate analysis of the music and participant data of the *Hooked!* game that allows for findings to be interpreted in terms of listening expectations of the participants.

Section 8.2 presents the analysis itself, and compares the new features to a set of symbolic features. We find that our corpus-based audio features are able to explain a comparable amount of variance to symbolic features. When the two types of features are used together, they supplement each other profitably. Along the way, we discuss the newly gained insights into what makes music recognizable, as revealed by the *Hooked!* data.

8.1 SECOND-ORDER AUDIO FEATURES

We begin by introducing the notion of second-order features, and proposing a fully specified adaptation of this idea for audio descriptors.

8.1.1 *Second-Order Features*

Second-order features are derivative descriptors that reflect, for a particular feature, how an observed feature value relates to a reference corpus [132]. There are several motivations for the use of second-order features.

First, they help in quantifying similarity and relevance of documents in the context of information retrieval. In information retrieval from text data, a common document description paradigm based on word counts involves *weights* that depend on the frequency of each word in a large corpus. Uncommon words are typically given more weight, as for example in 'TF×IDF' weighting, where TF (for term frequency) is the frequency of each term in the document, and IDF (for inverse document frequency) relates to the number of documents in the corpus that contain the term. Feature weighting helps estimating the relevance of a document in a retrieval context, and has been used as such for a very long time. The text analysis field of latent semantic analysis (LSA) is concerned with this kinds of text description [162].

Another motivation for the use of corpus-relative features could be to make the resulting feature more interpretable. They help contextualize the values a feature can take. Is 18.25 a high number? Is it a common result? Or if the feature is multivariate: is this combination of values typical or atypical, or perhaps representative of a particular style?

Finally, we shall show in section 8.1.4 that second-order features can be more cognitively plausible descriptions than the features we have been using so far. By giving us a means to approximately quantify expectations, distinctiveness and recurrence (the importance of which has been discussed in chapter 7), second-order features can be particularly useful in the analysis of recognizability and hooks.

8.1.2 *Second-Order Symbolic Features*

Second-order features have been used in symbolic music analysis, both of the retrieval and the corpus analysis kind. Like in text mining,

many features use the notion of document frequency, e.g., the number of songs in a large corpus that contain a given pitch interval.

The FANTASTIC toolbox by Müllensiefen implements many of these features. For example, the mtcf.mean.log.DF feature represents the average document frequency of all melodic motives or 'm-types' in a given melody, given a melody corpus.

M-types, inspired by the concept of *types* (entries in a dictionary) used in computational linguistics, are short sequences of symbols encoding pitch intervals and duration ratios of neighboring notes. M-types relate closely to the musical concept of melodic motives.

There is a strong parallel between the M-type counts used in the FANTASTIC toolbox, and the audio bigram features proposed in chapter 5 of this thesis. Both encode the relative occurrence of pitches in a specific order. We now discuss how the notion of document frequencies, and second-order features in general, can be adapted and used with audio bigrams and other audio descriptors.

8.1.3 *Second-Order Audio Features*

A fundamental difference between symbolic and audio representations of music, is that symbolic representations represent music as a streams of discrete event (e.g., notes, chords), while digital audio represents continuous, uninterrupted signals. This also applies to features: symbolic features operate on countable collections of events. Audio representations, even if they are discrete time series, based on frequency-domain computations or otherwise measured over short windows, represent continuous, uncountable quantities. This makes it impossible to apply the same operations directly to both, and alternatives must be found for the audio domain.

We define three types of second-order features. All represent how typical an observation is in a certain reference corpus. In statistical terms, the typicality of an observation in some feature space—how often does this value occur?—corresponds to the frequency density of this feature at the location of the observation. We distinguish between

one-dimensional descriptors such as loudness, and multivariate features such as audio bigrams.

Second-Order Audio Features in One Dimension

In one dimension, the most straightforward measure of typicality uses a density estimation method to estimate, for an observed feature value, the frequency density of the feature in the corpus. This approach of 'replacing feature values with densities' is also followed in the FANTASTIC toolbox.

Figure 31 shows, top left, a scatter plot of 100 simulated feature value observations (x-axis) and their associated frequency density (y-axis). The original 'first order' feature values are drawn from a standard -normal distribution ($\mu = 0$, $\sigma = 1$, $N = 100$). This has a particular downside, however: since by definition, more observations in the corpus will be associated with a higher feature density, the resulting distribution of the second-order feature will be heavily skewed towards higher values of typicality. The histogram on the top right of figure 31 shows the distribution of this second-order features based on density, for the same 100 simulated observations. This skew is a potential obstacle when the feature is to be used in statistical models, many of which assume normally distributed features, or at least minimally skewed distributions.

A typical method of dealing with this is to find a monotonous transformation that removes the skew from the distribution. Here, we propose a transformation based on *log odds*. Log odds are an alternative measure of probability. The log odds associated with the probability $p \in [0, 1]$ is given by the *logit* function:

$$\log \text{odds} = \text{logit}(p) \tag{72}$$

$$= \log(\frac{p}{1-p}) \tag{73}$$

Our log odds-based second-order feature, the 'logit ranked density' of a feature value x can formally be defined as the *log odds of observing a less extreme value in the reference corpus*. It is conceptually similar to a

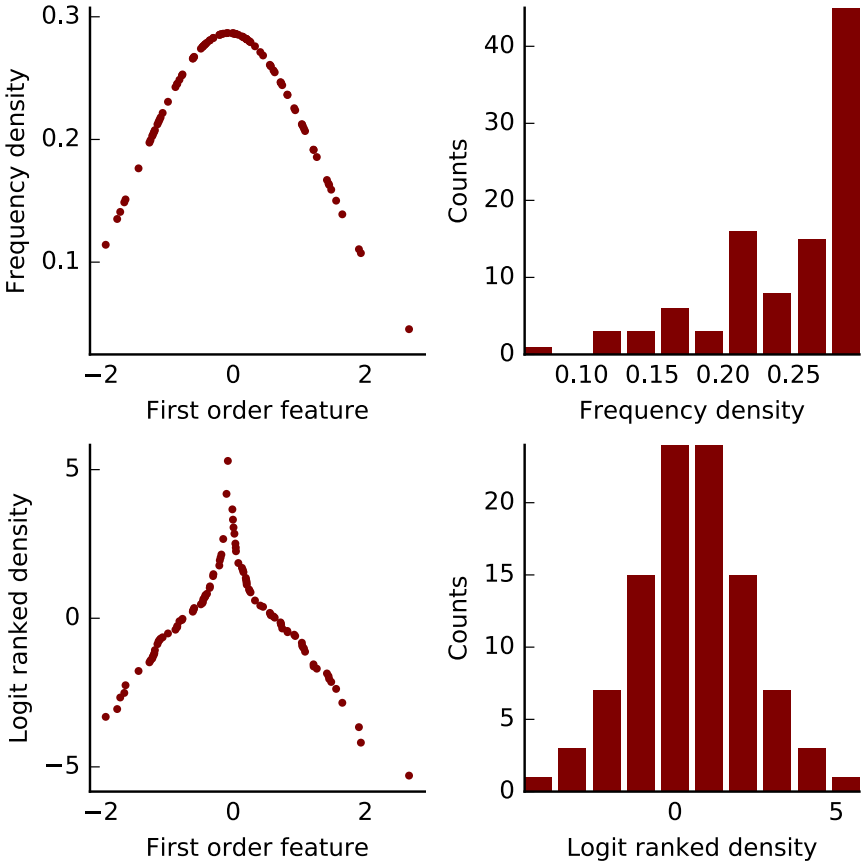


Figure 31.: Left: scatter plot of first order vs. second-order feature values for a sample of 100 simulated observations, and two second-order feature types: density (top) and logit ranked density (bottom). Right: histogram of second-order feature values for the same two feature types.

p-value, which measures the probability of observing a *more* extreme value, but we look at its complement, expressed as log odds.

We further propose a simple non-parametric approach to compute the above odds, based on ranking. By defining 'less extreme' as 'more probable', we can follow the density estimation approach described above to obtain probability density estimates $f(X)$ for all corpus values X and the observed feature value x. We then sort both $f(X)$ and $f(x)$ to find the rank of the feature value's density $f(x)$, and normalize it by the number of items in the corpus. Applying the logit function gives us the *logit ranked density*, hereafter, Z:

$$Z(X) = \text{logit}\left[\frac{\text{rank}\left(f(X)\right) - 0.5}{N}\right] \tag{74}$$

where N is the size of the reference corpus.

The lower half of figure 31 shows, on the left, a scatter plot for 100 simulated feature values (x-axis) and their second-order logit ranked density Z (y-axis). The distribution is somewhat unstable around the $x = 0$, but it is not divergent: as it is based on ranking, the distribution of $Z(X)$ is always bound to $\max(Z) = \text{logit}((N - 0.5)/N) = \log(2N - 1)$.

When we look at the distribution of $Z(X)$, shown on the right, we see that it is perfectly symmetrical. Indeed, again because of its definition based on ranks, Z always follows the same logistic distribution, which is bell-shaped, and generally very similar to a normal distribution. The feature can therefore be used out of the box for a variety of statistical applications.

If the first order feature X is one-dimensional, some form of density estimation is typically possible even if few data are available. Some caution is warranted when using Z where there are a limited number of observations, compared to the number of dimensions. The difficulty of multidimensional density estimation is widely acknowledged in statistics. In short, when the dimensions of a multidimensional feature are expected to be correlated (as is the case for chroma features), a covariance matrix must typically be estimated as part of any kind of density estimation. This increases the number of parameters to be estimated, and therefore the number of required data points—which may not always be available. In the FANTASTIC toolbox, too, "densities are

only computed for one-dimensional features because of the additional conceptual complexity and the high computational resources needed to estimate densities" [132].

Second-Order Audio Features in d Dimensions

For features with more than two or three dimensions, we now propose three methods of computing an alternative second-order feature. The first alternative is very simple: when a multivariate features has relatively independent dimensions by design (e.g., MFCC features), each dimension may be treated as a one-dimensional feature, and a meaningful Z based on density estimation can still be obtained. In practice this amounts to following equation 74 for Z above, but using a diagonal covariance matrix in the density estimation step.

The audio bigram features *MIB* and *HIC* (chapter 5) have 144 dimensions. We may be able to treat these as independent dimensions and get a useful estimate of typicality, however, since the audio bigram features can generally be understood as a probability distribution themselves, other measures of typicality may be more relevant. As a balanced compromise between a range of different options, we adopt two complementary measures of which the distributions are well-behaved.

The first approach is to compute *information* (I), an information-theoretic measure of *unexpectedness*. This measure assumes that the multidimensional first order feature itself can be seen as a frequency distribution F over possible observations in an audio excerpt (cf. term frequencies), and that a similar distribution F_{corpus} can be found for the full reference corpus. We define the $I(F)$ as the average $-\log F_{\text{corpus}}$ weighted by F:

$$I(F) \;=\; -\sum_{i=1}^{d} F(i)\ \log F_{\text{corpus}}(i) \tag{75}$$

The assumptions hold for *MIB*, *HIC* and *HI*, and produce well-behaved second-order feature values. The result is similar to *mean.log.TFDF*, *mtcf.mean.log.DF* and *mtcf.mean.entropy* in the FANTAS-

TIC toolbox. Information is also used as a measure of surprise, or prediction error, by Pearce and others in computational models of (music) cognition [53, 153].

The second measure, a measure of *expectedness* also used in the FANTASTIC toolbox, is a pragmatic, non-parametric measure of similarity between two vectors: Kendall's rank-based correlation τ, computed for the 'term frequencies' F and 'document frequencies' F_{corpus}. Kendall's τ counts the difference in the number of concordant and discordant pairs when the two vectors are sorted and the ranks are compared for each dimension of F. Both $I(F)$ and τ can be computed even for a small reference corpus.

8.1.4 *Song- vs. Corpus-based Second-order Features*

We can expand the notion of second-order features once more when we have access to a corpus of song *sections* rather than songs, as is the case for the Hooked data. Specifically, when a first-order description is available for several sections per song, we can define two reference corpora for every section: the large reference corpus, containing many sections from many songs, and a small, local reference corpus consisting only of sections from the same song. This in turn allows for two types of second-order features: corpus-based and song-based second-order features. In this section, we discuss the advantages of both types of features.

In a statistical learning perspective, expectations arise from statistical inference by the listener, who draws on a lifetime of listening experiences to assess whether a particular stimulus is to be expected or not. In chapter 7, we introduced Huron's three types of musical expectation, including *schematic* expectations, analogous to episodic and semantic memory, and veridical expectations. In short, schematic expectations arise from the 'auditory generalizations' that help us deal with novel, but broadly familiar situations. Veridical expectations are due to familiarity with a specific musical work. Finally, Huron also describes 'adaptive' expectations, which arise dynamically, upon lis-

tening. A repeated motive in a song you never heard before would generate this kind of adaptive expectations.

As the statistical learning paradigm goes, patterns that are more representative of the listener's listening history are more expected. The corpus-based second order features defined above measure typicality and surprise using a large corpus to approximate listening history. Therefore, we can use them to incorporate a crude approximation of schematic expectation in our analysis of hooks, or any other corpus analysis in which expectation plays a role. In the following section, we will refer to corpus-based second-order features as *conventionality*.

The song-based second-order features, by choosing as the reference corpus the set of all segments belonging to the same song, can be said to approximate 'local' expectations. Whether these are more adaptive or more veridical in nature is not entirely obvious—perhaps song-based second-order features cannot distinguish between the expectedness or surprise in some part of an unknown song as it comes along, and the expectedness of a song fragment to a listener who is familiar with the song (veridical expectations). In either interpretation, the features indicate how representative a segment is for the song, and to some extent, how much a segment is repeated. We will therefore also refer to song-based second-order features as *recurrence*.

8.2 DISCOVERY-DRIVEN HOOK ANALYSIS

Now that we have introduced second-order features, we can analyze our corpus of hooks. Using the data from the *Hooked!* experiment, we address the questions:

1. which attributes of the music, as measured by first and second-order features, predict the recognizability of sections of popular music?

This question will be approached by modeling differences between song sections of the same song, as will be argued later in this section.

Additionally, we consider this experiment a good test case to evaluate the newly proposed second-order descriptors described above. Therefore, we would like to know:

2. how do the proposed audio features behave and what aspects of the music do they model?

3. how much insight do audio-based corpus analysis tools add when compared to a symbolic feature set?

We first discuss data and features in sections 8.2.1—8.2.3, before introducing the statistical modeling approach in section 8.2.4.

8.2.1 *Data*

The Hooked experiment and its first implementation, *Hooked!* were described in chapter 7. The experiment tested how quickly and accurately participants could recognize different segments from each song in a collection.

For each song segment and each participant, the *Hooked!* data include a *recognition time* r. The recognition times of all trials in which a user knew the song fragment were combined into a *drift rate*, a single estimate of its recognizability roughly equal to the the reciprocal of the amount of time it would take a median participant to recognize the segment. Stimulus drift rates are commonly used as a measure of recognizability in timed recognition tasks.

The *drift-diffusion model* of memory retrieval, by Ratcliff, was the first cognitive model to propose such a measure [164]. The model assumes that, in a memory retrieval task with two possible answers, responses are driven by evidence accumulating over time, in a way that can be modeled by a continuous random walk process. Figure 32 shows a representation of this process. Here, time is shown on the x-axis, and the (non-monotonic) accumulation of evidence is shown on the y-axis. The random walk begins at a bias level z at a drift rate with mean u and variance s^2. A positive response ('I know this song!') is reached when the evidence hits the top match boundary a. A negative

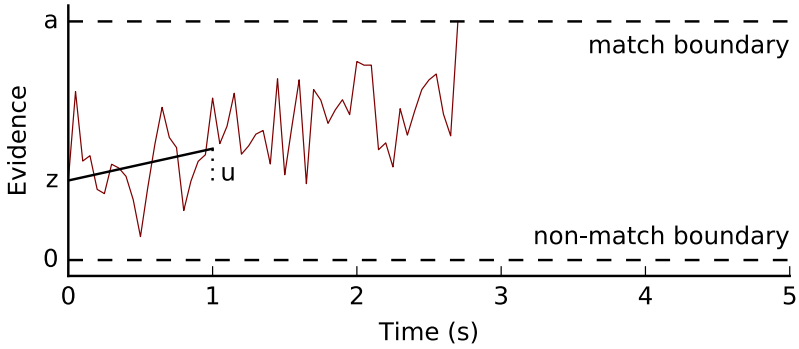


Figure 32.: Diagram of the Ratcliff drift diffusion model of memory retrieval [164]. The decision process is modeled as a random walk, beginning at a bias level z and ending when one of the the match boundaries 0 and a is reached. The mean rate of the random walk is the drift rate u.

response is reached when the accumulated evidence dives below the non-match boundary (x-axis).

The estimation of drift rates in the *Hooked!* dataset is based on a simplified, linear version of Ratcliff's model: the linear ballistic accumulator (LBA) [20]. Linear ballistic accumulator models are easier to fit to data than Ratcliff's original stochastic model. The LBA model, shown in figure 33, associates with each possible response a different 'accumulator', each with their own drift rate distribution (normal with mean v_i and variance s) and bias distribution (uniform between 0 and A). A response is reached when one the accumulators reaches the common match boundary b. LBA allows for more than two accumulators, so it can be applied to tasks with more than two possible responses.

This allowed us to adapt the LBA model and include three accumulators: one for trials in which a participant didn't know the song, one for trials in which the verification question was answered correctly,

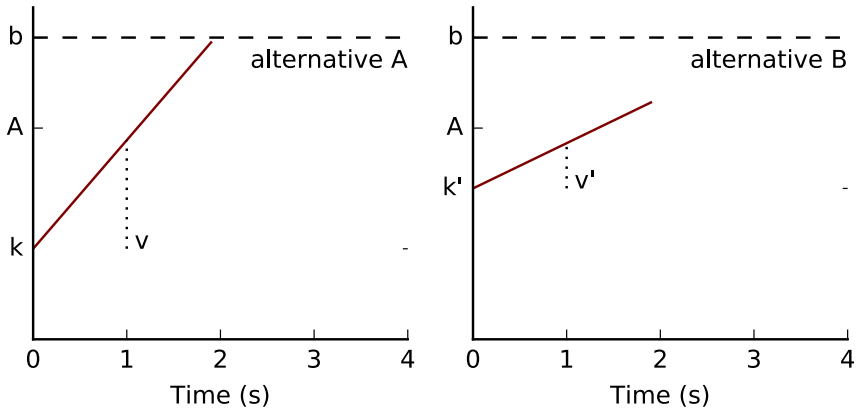


Figure 33.: Diagram of the linear ballistic accumulator model of memory retrieval [20]. The decision process is modeled by one 'accumulator' for each alternative response, each with their own drift rate distribution (normal with mean v_i and variance s) and bias distribution (uniform between 0 and A). A response is reached when one the accumulators reaches the common match boundary b.

				Dispersion		
		2nd-order			2nd-order	
	1st-order	Corpus	Song	1st-order	Corpus	Song
loudness	mean	Z(mean)	Z(mean)	std. dev.	Z(std.dev.)	Z(std.dev.)
sharpness	mean	Z(mean)	Z(mean)			
roughness	mean	Z(mean)	Z(mean)			
MFCC		Z(mean)	Z(mean)	total var.	Z(tot.var.)	Z(tot.var.)
pitch	mean	Z(mean)	Z(mean)	std. dev.	Z(std.dev.)	Z(std.dev.)
MIB		I, τ	I, τ	entropy	Z(entropy)	Z(entropy)
HIC		I, τ	I, τ	entropy	Z(entropy)	Z(entropy)
HI		I, τ	I, τ	entropy	Z(entropy)	Z(entropy)

Table 5.: Overview of the audio feature set used in the *Hooked!* data analysis.

and one for trials in which the participant failed the verification question [25]. In other words, we fit three drift rates per stimulus. All other parameters were set to depend on the participant (e.g., bias), or fixed.

To ensure a reliable fit, we iteratively excluded all song segments with fewer than 15 responses, and participants with fewer than 15 trials. We further excluded all segments from songs with fewer than 3 segments left. After these exclusions, 1715 song segments remained, taken from 321 different songs, representing data from 973 participants. An additional subset was created from 99 songs (536 segments) for which we were able to obtain symbolic transcriptions of the melody and bass line. This subset was used for the symbolic feature model, and to compare audio and symbolic features.

8.2.2 *Audio Features*

Two sets of audio descriptors were combined: first- and second order timbre descriptors, and first- and second-order pitch (melody and

harmony) descriptors. The total number of features is 44. All features were computed over 15-s segments starting from the beginning of each segment, as participants in the experiment were given a maximum of 15 s for recognition.

For timbre description, we used a feature set that is largely the same as the one used in chapter 4. Specifically, we computed the loudness (mean and standard deviations) for each segment, mean sharpness and roughness, and the total variance of the MFCC features. Instead of the pitch centroid feature, we obtained an estimate of pitch height using the *Melodia* melody extraction algorithm and computed the mean and standard deviation.[1]

For each of these one-dimensional features, we then computed the corpus-based and song-based second-order features Z as described in section 8.1.3 using a Python implementation.[2] Finally, we added song and corpus-based $Z(X)$ features based on the mean of the first 13 MFCC components. First-order features based on the MFCC means were not included because of their limited interpretability. An overview of the audio feature set is given in table 5.

For melody and harmony description, we used three of the features described in chapter 5: Melodic Interval Bigrams (*MIB*), Harmonic Interval Co-occurrence (*HIC*) and Harmonization Intervals (*HI*). HPCP were used as chroma features.[3] From these descriptors, we compute the entropy H as a first-order measure of dispersion.[4]

$$H = \sum_{i_1} \sum_{i_2} F(i_1, i_2) \log F(i_1, i_2) \tag{76}$$

The entropies were normalized as follows:

$$H' = \log \frac{H_{\max} - H}{H_{\max}} \tag{77}$$

As second-order features, the information I, and Kendall's τ were computed, as proposed in section 8.1.3.

1 `http://mtg.upf.edu/technologies/melodia`

2 code will be made available at `http://github.com/jvbalen`

3 `http://mtg.upf.edu/technologies/hpcp`

4 To capture as much variance as possible, entropy computation was performed on triads and trigrams before converting them to interval profiles.

8.2.3 *Symbolic Features*

For the symbolic reference feature set, we used a subset of 19 first-order and 5 second-order features from the FANTASTIC toolbox , computed for both melodies and bass lines. Second-order features were computed with both the song and the full dataset as a reference, yielding a total of 58 symbolic descriptors. Table 6 lists all features, with a short description. For exact definitions, see [132].

8.2.4 *Statistical Analysis*

There are two main particularities about the statistical analysis method that will be used in the analysis of the *Hooked!* data: first, it is a discovery-driven analysis, and second, it will be restricted to the analysis of within-song differences. We will now explain what both of these things mean.

Principal Component Analysis

Which attributes of music predict recognizability? Answering the question raised at the beginning of this section calls for a discovery-driven analysis method. This approach to corpus analysis is one of three types of research questions identified in section 3.5.1. It is an exploratory approach in which no particular hypothesis is tested. Typically, we are interested to know which of a candidate set of features correlates with a particular variable of interest. Examples are the approach followed in the analysis of choruses in chapter 4, Leman's analysis of audio features that predict walking speed, and Müllensiefen's analysis of oldness ratings in a melodic memory task [104, 134].

A challenge that arises with this approach, one of several reviewed in chapter 3, is that it typically requires many tests to assess the correlation of several feature with the variable of interest. As a result, a sound strategy is needed to minimize *false positives*, discoveries due to chance. In the three examples above, three strategies are followed: in chapter 4, a probabilistic graphical model is learned, with significance

first-order feature	description
d.median	median note duration
d.range	note duration range (maximum − minimum)
d.entropy	entropy of the note durations distribution
p.std	standard deviation of the pith distribution
p.range	pitch range
p.entropy	entropy of the pitch distribution
i.abs.mean	mean absolute pitch interval
i.abs.std	standard deviation of absolute pitch interval
i.abs.range	absolute pitch interval range
i.entropy	entropy of the pitch interval distribution
len	length of the melody in notes
glob.duration	global duration of the melody
note.dens	number of notes per second
int.cont.grad.mean	mean gradient of the pitch contour
int.cont.grad.std	standard deviation of the gradient of the contour
tonalness	highest of 24 correlations with Krumhansl's key profiles
tonal.clarity	ratio of highest and second-highest key correlation
mean.entropy	mean entropy of the distributions of length-n m-types
mean.productivity	mean of the fraction of length-n m-types appearing only once in the melody
second-order feature	**description**
mtcf.mean.log.DF	mean document frequency of the melody's m-types
mtcf.mean.log.TFDF	TF-weighted mean document frequency of the melody's m-types
mtcf.mean.productivity	mean of the fraction of length-n m-types appearing only once in the corpus
mtcf.TFIDF.m.entropy	entropy of TF-IDF weights of the melody's m-types
mtcf.TFDF.kendall	Kendall τ for TF and DF of the melody's m-types

Table 6.: List of symbolic features used in the *Hooked!* data analysis. All features from the FANTASTIC toolbox [132].

levels for each of the tests adjusted based on the total number of tests involved. In [104], a set of linear models is fit, and cross validation is used to perform model selection on the results. In [134], partial least squares regression (PLSR) is used to combine features into components before fitting a linear model.

In a simpler variation on the PLSR approach by Müllensiefen, we will use principal component analysis (PCA) before fitting the features to the drift rates, as a way of identifying groups of features that may measure a single underlying source of variance. PCA reduces the feature space to a more manageable number of decorrelated variables. This reduces the number of tests required in the next stap of the analysis, a linear model, and thereby the risk of false positive discoveries.

PCA was applied to both the audio and symbolic features, separately. Features were centered and normalized before PCA, and the resulting components were transformed with a varimax rotation to improve interpretability. This orthogonal transformation of the principal components finds rotations in which components have just a few highly-loading parameters, and variables load onto just a few of the components. We selected the number of components to retain (12 in both cases) using parallel analysis, a heuristic method that identifies the number of components needed to model most of the information in the data, by comparing the ranked principal components to those of a randomly sampled dataset with the same number of variables and observations [72]. Not all 12 components representing the symbolic features will be discussed here, but they were considered coherent and interpretable enough to proceed with the analysis. The audio feature components will be discussed as part of the results.

Linear Mixed Effects Model

The second main idea behind our approach to statistical analysis, is that we want to exploit the structure of the *Hooked!* dataset: as we have drift rate estimates for each of the songs sections, we can perform an analysis that looks only at differences between sections of the same song. This allows us to ignore between-song variation, a component

of recognizability that may be dominated by the effects of a variety of extramusical factors, e.g., difference in age of the song, marketing, radio play or social appeal. Instead, we focus on within-song variation, which is much more related to our definition of hooks as the most recognizable part of the song, regardless of its absolute 'catchiness'.

We use a linear mixed-effects regression (LMER) model to fit the feature principal components to the drift rates. Mixed-effects models can handle 'repeated-measures' data where several data points are linked to the same song and therefore have a correlated error structure. The *Hooked!* data provide drift rates for individual sections within songs, and one would indeed expect considerably less variation in drift rates within songs than between them: some pop songs are thought to be much catchier than others overall. Linear mixed-effects models have the further advantage that they are easy to interpret due to the linearity and additivity of the effects of the predictor variables. More complex machine-learning schemes might be able to explain more variance and make more precise predictions for the dependent variable, but this usually comes at the cost of the interpretability of the model.

We fit three models: two including audio components only, one including symbolic components only, and one including both feature types, and used a stepwise selection procedure at $\alpha = 0.005$ to identify the most significant predictors in each model. The audio-only model is fit twice to facilitate comparison between audio and symbolic features: once using the full set of 321 songs and again using just the 99 songs with transcriptions.

In all models, the dependent variable was the *log* drift rate of a song segment and the repeated measures (random effects) are handled as a random intercept, i.e., we add a per-song offset to a traditional linear regression (fixed effects) on song segments, with the assumption that these offsets be distributed normally:

$$\log v_{ij} = \boldsymbol{\beta}\mathbf{x}_{ij} + u_{i0} + \epsilon_{ij} \tag{78}$$

where i indexes songs, j indexes segments within songs, v_{ij} is the drift rate for song segment ij, $\mathbf{x}_{ij}$ is the vector of standardized feature

component scores for song segment ij plus an intercept term, the $y_i \sim N(0, \sigma^2_{\text{song}})$, and the $\epsilon_{ij} \sim N(0, \sigma^2_{\text{residual}})$.

8.3 RESULTS AND DISCUSSION

8.3.1 *Audio Components*

The results of the principal components analysis of the audio features set, the component loadings, are shown in a table in appendix A. The component loadings (correlation coefficients between the extracted components and the original features) tell a consistent story. The first 11 components break the audio feature set down into three timbre components (first order, conventionality, and recurrence) and three entropy components (idem), two features grouping conventionality and recurrence for melody and harmony, respectively, and three more detailed timbre components correlating with sharpness, pitch range and dynamic range.

The last component (component 9 in the table in appendix A) is the most difficult to interpret. It is characterized by an increased dynamic range and MFCC variance, and a typical pitch height. We hypothesize that this component correlates with the presence and prominence of vocals. It is reasonable to assume that the most typical registers for the melodies in a pop corpus would be the registers of the singing voice, and vocal entries could also be expected to modulate a section's timbre and loudness. This hypothesis is also consistent with our own observations while listening to a selection of fragments at various points along the component 9 scale. The high end includes a number of *a capella* or minimally accompanied vocal segments along with a few prominent guitar solos in the vocal register. The verse section from Alicia Keys's No One (2007) is a representative example, with very prominent vocals and only sparse accompaniment. The low end consists primarily of instrumental breaks with relatively undefined melodic content or segments with notably faded vocals, as in the instrumental break of Foo Fighters' Everlong (1997).

Overall, the neatness of the above reduction attests to the advantage of using interpretable features, and to the potential of this particular feature set. Specifically, the tendency of the components to distinguish between conventionality and recurrence suggest that the distinction between song-based and corpus-based second-order features is indeed informative.

8.3.2 *Recognisability Predictors*

Results

Table 7 contains the results of all four linear mixed effects models, showing the fixed effect coefficients, the random intercepts and R^2 values for each model. As expected, the random intercepts per song explain a large amount of variance in the drift rates: between 37 and 40%. However, we are mostly interested in the within-song differences. The coefficients for these fixed effects can roughly be interpreted as percent increase in drift rate per unit of standard deviation in the component (because the dependent variable is logarithmically scaled and the correlation coefficients are relatively low), which makes interpretation easier.

A look at the first column of results for the linear mixed effects model confirms that the audio features are indeed meaningful descriptors for this corpus. Eight components correlate significantly, most of them relating to conventionality of features. This suggests a general pattern in which more recognizable sections have a more typical, expected sound. Another component, timbral recurrence, points to the role of repetition: sections that are more representative of a song are more recognizable. Finally, the component with the strongest effect is Vocal Prominence.

The model based on symbolic data only, in the third column, has just two components. This is possibly due to the reduced number of sections available for fitting. The results, in the second column, for an audio-based model fit on the reduced dataset of 99 songs supports this explanation, as it also yields just two components. In the presence

Parameter	Audio[a] $\hat{\beta}$	Audio[a] 99.5 % CI	Audio[b] $\hat{\beta}$	Audio[b] 99.5 % CI	Symbolic[b] $\hat{\beta}$	Symbolic[b] 99.5 % CI	Combined[b] $\hat{\beta}$	Combined[b] 99.5 % CI
Fixed effects								
Intercept	−0.84	[−0.91, −0.77]	−0.67	[−0.78, −0.56]	−0.62	[−0.73, −0.51]	−0.63	[−0.74, −0.53]
Audio								
Vocal Prominence	0.14	[0.10, 0.18]	0.11	[0.04, 0.17]			0.08	[0.01, 0.15]
Timbral Conventionality	0.09	[0.05, 0.13]						
Melodic Conventionality	0.06	[0.02, 0.11]						
M/H Entropy Conventionality	0.06	[0.02, 0.10]						
Sharpness Conventionality	0.05	[0.02, 0.09]						
Harmonic Conventionality	0.05	[0.01, 0.10]						
Timbral Recurrence	0.05	[0.02, 0.08]						
Mel. Range Conventionality	0.05	[0.01, 0.08]	0.07	[0.02, 0.13]			0.07	[0.01, 0.12]
Symbolic								
Melodic Repetitivity					0.12	[0.06, 0.19]	0.11	[0.05, 0.17]
Mel./Bass Conventionality					0.07	[0.01, 0.13]	0.08	[0.01, 0.14]
Random effects								
$\hat{\sigma}_{\text{song}}$	0.39	[0.34, 0.45]	0.35	[0.26, 0.45]	0.34	[0.25, 0.44]	0.32	[0.24, 0.42]
$\hat{\sigma}_{\text{residual}}$	0.48	[0.45, 0.50]	0.40	[0.37, 0.44]	0.39	[0.35, 0.43]	0.38	[0.34, 0.42]
R^2_{marginal}[c]	.10		.06		.07		.10	
$R^2_{\text{conditional}}$[c]	.47		.46		.47		.47	
$-2 \times$ log likelihood	2765.61		699.81		576.74		558.11	

Note. Random-intercept models, grouping by song, for the given feature types after step-wise selection using Satterthwaite-adjusted F-tests at $\alpha = .005$. Component scores were standardized prior to regression.

[a] Complete set of 321 songs ($N = 1715$ segments).

[b] Reduced set of 99 songs with symbolic transcriptions ($N = 536$ segments).

[c] Coefficients of determination following Nakagawa and Schielzeth [138].The marginal and conditional coefficients reflect, respectively, the proportion of variance in the data that is explained by the fixed effects alone and the proportion explained by the complete model (fixed and random effects together).

Table 7.: Estimated coefficients and variances for audio and symbolic components predicting the relative recognizability of popular song segments.

of less data, only the most important components stand out. The symbolic and reduced audio model seem to be comparable in power with a marginal R^2 of 0.06 and 0.07, respectively (see figure caption for definitions).

The top symbolic features that make up the first of the significant components are melodic entropy and productivity, both negatively correlated, suggesting that recognizable melodies are more repetitive. The top features that make up the second components are *mtcf.mean.log.DF*, for the melody (song-based and corpus-based), and negative *mtcf.mean.productivity* (song-based and corpus-based for both bass and melody). This suggests that recognizable melodies contain more typical motives (higher codument frequencies, lower second-order productivity).

The last column shows how the combined model, in which both audio and symbolic components were used, retains the same audio and symbolic components that make up the previous two models. The feature sets are, in other words, complementary: not only are all four components still predictive at $\alpha < 0.005$, the marginal R^2 now reaches 0.10, as opposed to 0.06 and 0.07 for the individual models. This answers the last of the questions stated in section 8.2: for the data in this study, the audio-based corpus analysis tools contribute substantial insight, and make an excellent addition to the symbolic feature set.

Discussion

We briefly discuss our findings on the properties of hooks. First, the presence of vocals appears to be the strongest predictor of recognizability. We see this as an unsurprising but important result: it suggests that vocal melodies are very important in the recognition of popular music.

Second, sections that are more conventional in terms of melody, harmony, bass and timbre are more recognizable: 7 conventionality components in total, across both the audio and the symbolic model, show a consistent positive correlation between conventionality and recog-

nizability. In other words, if there were to exist a positive effect of *distinctiveness* on recognizability, as suggested in some of the hook hypotheses in chapter 7, no evidence is found for it in this analysis. The analysis suggests rather the opposite: recognizable sections are more typical than they stand out.

Finally, the data suggest that recognizable song sections are more repetitive (as measured by symbolic melody repetitiveness), and more repeated (as measured by timbre recurrence). This does align with some of the related findings reviewed in chapter 7: repeated exposure and recognizability go hand in hand—at least, as measured using these two sets of features.

8.4 CONCLUSIONS AND FUTURE WORK

In this chapter, we have presented a new approach to corpus-level audio description, and a new discovery-driven analysis of popular music hooks. We introduced three general-purpose second-order audio descriptors: the 'logit ranked density' Z, information I and Kendall's τ, and the notion of song-based and corpus-based second-order features. In the hook discovery experiment, two features sets were compiled: an audio feature set based on the new the audio description methods and a symbolic reference feature set. We then used PCA and LMER to predict, from these features, recognizability of song fragments in the *Hooked!* dataset.

From the results and discussion of the statistical analysis we conclude that the harmony and melody descriptors, the corpus-based second-order features and the song-based second-order features contribute new and relevant layers of information to the corpus description. From the results of the audio analysis, we conclude that sections with vocals and sections that are most representative of the song in terms of timbre, are better recognized. Recognizable song sections also have a more typical, expected sound. From the symbolic results, we conclude that recognizable melodies are more repetitive, and contain less atypical motives. In short: vocals, conventionality and repetition best predict recognizability.

Finally, we conclude that an audio corpus analysis as proposed in this paper can indeed complement symbolic corpus analysis, as the experiment sees both kinds of features explaining an important share of the variance in the data. This opens up a range of opportunities for future work. These will described in the last chapter of this thesis.

9

CONCLUSIONS

9.1 CONTRIBUTIONS

The goal of this thesis, as stated at the beginning of chapter 1, is to make a number of contributions to the scientific study of music based on audio corpus analysis. We set out to address three sets of research problems in particular. The first problem is that there is little information on what makes an audio descriptor a good descriptor for corpus analysis research, and that more such adequate descriptors may have to be developed. The second problem is that audio corpus analysis methods themselves, too, haven't been charted and may need to be improved. Third, we addressed two goals that are central to the COGITCH project. The first is a lack of audio description techniques and similarity models for retrieval and research on musical heritage collections. The second is the ambition to gain insight on the notion of 'hooks'. We now review the contributions made to address each of these problems.

9.1.1 *Audio Description*

The first set of contributions has been to map and extend the pool of adequate audio description techniques that are available for audio corpus analysis research. To this end, we began chapter 2 with a review of existing audio features for the description of melody, harmony and musical timbre. In chapter 3, based on a review of the corpus analysis literature, a list of guidelines was deduced to guide the choice

of audio features for corpus analysis research, related to robustness, dimensionality and interpretability. In chapter 4, we then presented a set of features that can be used to describe popular song sections, including psychoacoustic features, and simple harmony and melody descriptors.

In chapters 5 and 6, we introduced 'audio bigrams', a new family of multidimensional harmony and melody descriptors. They were defined in chapter 6 as measuring the co-occurrence of salient pitch events. Six examples were introduced in chapter 5. They are inspired by the notion of bigrams and trigrams in text and symbolic music analysis. Mathematically, all six relate to distributions over pairs of melodic and harmonic pitches and pitch intervals that occur close together in an audio excerpt.

9.1.2 *Audio Corpus Analysis*

The second set of contributions has been to review and extend the available methods for audio corpus analysis. First of all, this called for a review of the disciplines and interdisciplinary research contexts in which empirical music research takes place, allowing us to position audio corpus analysis research among them (chapter 1). In chapter 3, we then reviewed the most important work done in corpus analysis research. This resulted in a set of guidelines for future choices of research questions, data, audio descriptors and analysis methods in section 3.5.

In chapter 4, we presented the first use of a probabilistic graphical model in the analysis of audio features. We showed that it is possible to study the relation between a variable of interest—chorusness—and a selection of reliable audio features while controlling for confounding correlations between the audio features.

Expanding on the features evaluated in chapters 4 and 5, and inspired by methods from latent semantic analysis and symbolic corpus analysis, we also proposed the first 'second order' or corpus-relative audio features (chapter 8), quantifying the distinctiveness and recurrence of audio feature values in a corpus. In a corpus analysis of song

sections and hook annotations, a selection of first and second-order audio features were shown to be both competitive and complementary to symbolic first- and second-order features.

Finally, in the same corpus study of hooks, we also introduced two methods for the statistical modeling of within-song variation. The notion of song-based second-order features can quantify distinctiveness and recurrence within a song. And a statistical model of song sections can quantify differences between song sections within songs, while controlling for differences between the songs themselves.

9.1.3 *Music Similarity and Hooks*

The third set of contributions pertains to the goals of the COGITCH project: improving audio similarity models for music heritage collections, and uncovering the properties of hooks.

Music Similarity

The challenges of modeling music similarity at scale were explained in chapter 1, contrasting alignment and non-alignment-based solutions to cover song detection. In chapters 5 and 6, we have presented several new pitch description methods. Throughout both chapters, the proposed description techniques were evaluated by applying them to the song similarity problem of cover detection.

The six novel pitch descriptors proposed in chapter 5 are all fixed-size representations of pitch use in a song or song segment. This makes them a useful asset in the design of efficient music similarity models. Three features were evaluated in a series of cover song experiments on two datasets: a dataset of early to mid-20th century translated songs from the *S&V* record collection, and a dataset of more recent cover songs used for benchmarking performance. Results showed a reasonable performance on either task when compared to other scalable systems.

Chapter 6 shows how audio bigrams relate to a range of existing 'soft audio fingerprinting' algorithms, algorithms for content-based

identification of music recordings. We also showed that in its most general formulation, the computation of audio bigram features can be fully vectorized, and even formulated entirely in terms of neural network components (convolutions, matrix multiplications and non-linearities, such as rectified linear units) making highly efficient implementations possible. Finally, we have presented PYTCH, an implementation of the audio bigram features paradigm in Python, for use in audio description, song similarity models and retrieval.

Analysis of Hooks

The second project goal was to use new audio description and corpus analysis methods to gain insight into the phenomenon of catchiness and hooks.

The analysis of choruses in chapter 4 served as an experiment to prepare for this eventual goal. Choruses are a recurring object of study in music information retrieval, and are often said to have a catchy and memorable quality. By studying choruses, we could use the readily available datasets used for structure analysis and segmentation, to get a first insight into what makes a piece of music catchy. The analysis showed that choruses in the Billboard charts are perceptually sharper and rougher than other sections. They also have a smaller dynamic range and greater variety of timbre. Finally, choruses feature a higher and more salient pitch, a trend that is already present in choruses of songs from the first half of the twentieth century.

For a deeper understanding of the properties of hooks, a definition of hooks was first required. In chapter 7, we introduced the notion of catchiness, and defined hooks as the part of a song that is most recognizable. After reviewing the prevailing hypotheses on what makes music catchy, we then described Hooked, a game we made to collect a dataset of hooks. The analysis of the music and recognizability estimates gathered using Hooked followed in chapter 8. It involved several of the audio descriptors introduced in chapters 4–6, and the novel concept of corpus-based and song-based second order features discussed above. A principal component analysis of the selected fea-

tures revealed twelve interpretable dimensions that could be used to match the audio features to recognizability. The results, controlling for differences between songs, show how sections with vocals and sections that were most representative of the song in terms of timbre, are generally more recognizable. Recognizable song sections also have a more typical, expected sound, as measured by several corpus-based second order features. In other words, hooks are characterized by the presence of vocals, and components that suggest repetition and conventionality. This is confirmed in an analysis of melodic transcriptions using similar, symbolic features: recognizable melodies are more repetitive and contain less atypical motives.

9.2 LOOKING BACK

Having reviewed contributions, we can now look back and critically assess them in light of the goals set in chapter 1 (section 9.2.1) and the methodological guidelines formulated in chapter 3 (section 9.2.2).

9.2.1 *Research Goals*

Have the research goals set at the beginning of this thesis been reached? For the most part, we believe they have. The above discussion lists several of the new approaches to audio description that have been have been proposed and tested as part of this thesis. However, not all opportunities to leverage the full potential of MIR for corpus analysis, were seized. We give a brief overview.

First, *rhythm description* has largely been absent from this thesis. This is unfortunate—strategies for rhythm description would be a valuable extension of the corpus analysis tool set. As chapter 5.1.1 explains: at the level of distributions of basic patterns, rhythm description proves to be significantly more challenging than harmony and pitch description. In the context of polyphonic audio, rhythm perception relies on two inference processes that are notably hard to model: note onset detection (especially of non-percussive onsets) and streaming. A lot of music perception and signal processing work is

still to be done when it comes to modeling these two perceptual skills, more than we could have done as part of this thesis.

Second: melody, harmony and timbre, too, have more facets than can be measured by the audio bigram descriptors introduced in this thesis—e.g., we talked about melodies but not about melodic contours, we talked about harmony but not about different voicings. Many other description methods could have been implemented, tested or developed to further characterize melody and harmony for corpus analysis research but we decided to leave these for future work.

Similarly, some of the statistical analysis methods encountered in the literature review of corpus analysis studies have not been applied or evaluated. In the two sets of original corpus studies we presented (chorus analysis and hook analysis) we used hypothesis testing, graphical models and classification (chapter 4), and a linear mixed effects model with principal components analysis (chapter 8). Methods we have not yet explored include large feature set analyses with feature selection (e.g., using cross validation, as in Leman's study on walking speed [104]), or Bayesian models (as we are currently using in an ongoing analysis of the *Hooked on Music* data).

Finally, we acknowledge that our efforts to improve content-based similarity, which we re-framed as 'soft audio fingerprinting' in chapter 6, have not yet yielded the powerful solutions we aimed for in the COGITCH project. However, we have provided a new theoretical perspective on an array of soft audio fingerprinting approaches. The resulting audio bigram paradigm was formally defined and implemented, and is ready to be tested in applications like large-scale content-based matching.

In short, many of the goals were achieved. The work that we havent been able to address centers on four issues: (i) the continued lack of validated rhythm description methods, (ii) the many possibilities of extending our approaches to melody, harmony and timbre description, (iii) opportunities to apply several more statistical methods to corpus analysis, and (iv) the application of audio bigram-based features to large scale document matching in musical heritage collections.

9.2.2 *Methodology*

Have the contributions listed above given sufficient consideration to the methodological guidelines in chapter 3? The guidelines or 'desiderata' for a good audio corpus analysis strategy pertain to choices in research questions, data, descriptors, analysis methods.

Have we used only robust, low-dimensional and informative features, as prescribed? Our pitch description features (chapter 5) mostly comply. They require a strategy for melody extraction, but not for note segmentation or other music transcription hurdles. They are rather high-dimensional, but they were complemented in chapter 8 with first and second-order aggregation functions that provide a useful one-dimensional summary of its dimensions. And they are informative: each of the dimensions can be interpreted as a probability of observing some combination of pitch classes or pitch class intervals.

In our choice of statistical analysis methods we have been cautious about false positives and overfitting. Significance levels were always set according to the number of tests in an experiment, and the number of parameters to each model was consistently kept in check with the amount of data, partially due to the aforementioned summarization of feature dimensions.

We have also devised, on two occasions, explicit strategies to control for confounding variables. In the analysis of choruses, we used a probabilistic graphical model—a statistical model of conditional independences rather than just correlations. This allowed us to identify different kinds of relationships between chorusness and audio features, even if those features were correlated. In the analysis of hooks, we presented a statistical analysis of song sections in which we used a mixed effects model to find trends in a corpus of song sections while controlling for differences between songs.

Finally, in each set of 'experiments' in part ii (chorus analysis and cover detection), two different datasets were used, which allowed us to corroborate the most important conclusions—not just different samples drawn from a larger dataset, but new, 'idiosyncratic' data.

A similar approach will be followed in the further analysis of hooks: chapter 8 presented the analysis of the *Hooked!* data, an analysis of the *Hooked on Music* data is still underway.

An example of a case in which, arguably, better choices could have been made, is in the choice of data. The Billboard dataset, used in chapter 4, has been very carefully sampled from the clearly defined population that is the Billboard charts. The Billboard charts themselves, however, were shown in chapter 3 to suffer from more biases and discontinuities than desirable for a supposedly authoritative metric of song popularity, mostly because of the way sales were measured by Billboard over the decades since the beginning of the charts. In general, popularity is an elusive concept which, many musicologists would argue, cannot be captured in a single number. The same music can be wildly popular in one place and unknown in another. And some songs are popular at the time of their release but hardly known today—many chart-topping songs from the past decades haven't made it into the collective memory.

Biases and inconsistencies therefore also exist in the Top 2000 list from which the Hooked data were sampled—possibly problematic, even if we control for some of the effects of inter-song differences: not all songs can be expected to have the same kind of hooks. Making an unbiased selection of well-known songs is a nearly impossible task, but it is worth reflecting on how the song collections could nonetheless have been more carefully sampled, as it is important for transparency of research results and, therefore, the integration of findings in the musicological discourse.

Finally, one could argue that we could have made a larger contribution by assessing the output of not just one, but several analysis methods per dataset, and comparing the findings. Each time a music corpus was analyzed, we have focused on showing that there exists a viable corpus analysis strategy that can be used to gain insight into the data.

However, in choosing a good audio description or a corpus analysis method, we are skeptical that simple, widely applicable answers exist. Each corpus analysis problem may call for new audio description

approaches or a different corpus analysis strategy. Conclusions about evaluation of descriptors and analysis methods might therefore not generalize to other contexts if they are based on a limited number of cases, e.g., the ones presented in this thesis. What makes our investigation valuable, then, are not the just the corpus studies and their results, but the literature review and methodological guidelines in chapter 3 and, above all, the description and analysis methods themselves. New methods were tested, but in the spirit of the new empirical method, not to benchmark them with respect to some measure of performance, but to show how they can be used and demonstrate their potential in one or two real-world music analysis problems.

9.3 LOOKING AHEAD

We now look ahead at the planned, upcoming work (section 9.3.1) and the most promising future work after that (section 9.3.2).

9.3.1 *Ongoing Work*

The CATCHY *Toolbox*

In the near future, we plan to release the code we used to compute second-order audio features as a small toolbox that can be used on top of PYTCH. It will be released on Github under the same name as the paper in which second-order audio features and the *Hooked!* analysis were introduced: CATCHY, or 'corpus analysis tools for computational hook discovery'.[1] The CATCHY toolbox will be tested together with colleagues at Goldsmiths University of London, as part of their research on the properties of earworms.

Hooked on Music *Data Analysis*

An important unfinished element of the COGITCH project is the analysis of the response times and accuracies gathered in the *Hook on Music*

1 `http://www.github.com/jvbalen/catchy`

game, the UK version of Hooked, as the data collection stage of this experiment has only recently closed. In the coming months, we will finalize the analysis of the participant data and the music.

Rather than using the exact same analysis strategy, we now aim to integrate the LBA model of memory retrieval into the mixed effects model. The result we aim for is a hierarchical Bayesian model in which all parameters can be estimated at once. Having access to the responses of up to 100 times more participants than were available in the *Hooked!* dataset will, hopefully, yield precise estimates of each of these parameters.

The results of the analysis will then be compared to the results obtained for the *Hooked!* data to see if our earlier findings are confirmed. Trends that are very significant but not shared may shed some light on the difference between the music in both datasets, and between the Dutch and UK music listeners.

9.3.2 *Future work*

Section 9.2 reviewed the contributions of this thesis in terms of the goals and the methodological perimeter set in part i. Here, we distill from this discussion three important and promising avenues for future work.

Rhythm Description

As said in above in section, the analysis of rhythm has been given very little attention in this thesis, even though we believe it should be part of a the corpus analysis toolbox. In future work, we believe rhythm description should be a priority.

Corpus studies such as those by Serrà et al., Mauch et al. (see section 3.4), and ourselves (chapter 8) paint a skewed picture of popular musing by measuring diversity, change, distinctiveness and repetition only in terms of melody, harmony and timbre. For example, claims that popular music has become increasingly homogeneous solely on the basis of those three musical facets ignore the possibility that har-

monic and melodic complexity might have been replaced with rhythmic complexity, e.g., due to the rise of hip hop and electronic music in the last decades. Even though hypothetical—we don't know how rhythmic complexity evolved—this example illustrates the potential impact of methodological decisions. First, the overall conclusion might have looked very different if rhythm was taken into account. Second, it illustrates how a mildly Euro-centric methodology (giving priority to melody, harmony and timbre over rhythm) can lead to conclusions with decidedly Euro-centric connotations (music based on loops, rap and sampling is somehow less complex). It reminds us to consider the concerns of new musicology at the end of last century: researchers bring their own cultural biases into the lab.

Audio Bigrams and Learned Fingerprints

In chapter 6 we proposed the umbrella task of soft audio fingerprinting. Essentially, soft audio fingerprinting is any kind of content identification in which a fixed-size representation is used to efficiently compare documents. At the end of the chapter, it was suggested that, given a specific soft audio fingerprinting problem (e.g., sample detection, remix detection or efficient cover song identification) and a ground truth of related documents, it may be possible to learn an optimal audio bigram representation of the music that is to be analyzed.

This approach has not been tested. However, we consider this a very promising path to efficient and versatile fingerprinting. The potential of feature learning techniques to outperform traditional information retrieval methods has already been shown in several other MIR tasks, so it can be expected that strategies exist in which feature learning works for fingerprinting applications as well. What makes a task like fingerprinting or cover detection difficult is that, unlike chord detection or tag prediction (see section 2.2), it cannot be trivially reduced to a typical classification problem. First, any given pair of remixes or cover songs tends to differ in length. Current feature learning systems tend to crop or scale their input data rather arbitrarily, in a way that would be problematic for a remix or cover song recognition system

(often, selecting a 30 second fragment at random). Second, the order in which the musical material appears is very important in cover detection, as the success of alignment-based systems suggests. And finally, training a classifier typically requires a rather large number of examples per class. In remix detection and cover detection, there are often only one or two cover versions of a song.

An audio bigram-based fingerprinting system, if implemented using the convolutional neural network components as described in section 6.3.3, gives us a way to deal with the first two problems: the matrix product reduces each song to a set of n $k \times k$ matrices (where n and k are predefined constants) that represent ordered occurrences of events. A possible strategy to overcome the third problem is to approach fingerprinting first as a binary classification problem, classifying pairs of documents as either the same or not the same (e.g., in a way similar to the recently proposed 'siamese' network architecture used in [163]). Then, having trained such a model, one of the learned, intermediate representations can be used as the basis of a fingerprint for matching at scale.

The current implementation of the audio bigram feature in the PYTCH toolbox (also presented in chapter 6) makes use of Numpy and Scipy, two Python toolboxes for numerical modeling that are well suited for vectorized computations, but not ideal for learning representations. Tools that are naturally suited for this problem are hybrid (symbolic and numerical algebra) toolboxes like Theano[2] and Tensorflow[3], which use techniques from symbolic computing to optimize the numerical manipulation of vectors, matrices and tensors. An adaptation of the audio bigram toolbox to work with one of these toolboxes could make it possible to test the potential of learning fingerprints from a dataset of examples.

In the longer term, we hope to be able to use this approach to improve efficient document matching, so that it can be used in musical heritage-related soft audio fingerprinting problems such as cover song

2 http://www.deeplearning.net/software/theano

3 http://www.tensorflow.org

detection in the S&V collection, and matching the MI's recordings of monophonic folk songs with S&V's digitized 78 RPM records.

The Future of Hooked

The data obtained with the Hooked games has already been put to good use. However, we believe much more research can be done, even after the first analysis of the *Hooked on Music* data is completed. In the analysis of each dataset, we have incorporated a notion of 'listening history' of the participants, which was used to compute distinctiveness of individual song fragments, and estimated using the whole of the music corpus used in each of the games.

A powerful extension of this approach could be pursued by modeling a more detailed listening history for each of the participants. Since we have data on how well they were able to recognize each of the stimuli, this can even be done fairly straightforwardly. The most straightforward approach would be to perform clustering of the users based on their response times per song. Advanced variants of this approach exist in music recommendation, specifically designed to deal with the sparsity of information that occurs when not all users listen to all music. The underlying idea is to factorize the matrix of users' response times into a matrix of taste profiles (each represented by a weighed subset of all the songs) and a matrix of participants listening preferences (representing each user's preferences as a weighted combination of taste profiles); see e.g, [73]. Integrating this kind of listener profiles may help in exposing interactions between distinctiveness and recognizability that our models can currently not observe.

One similar line of research has already been initiated by Burgoyne et al. [24]. In a 2015-2016 update to the *Hooked on Music* game, players are presented, as they progress, with songs they are increasingly likely to know. The updates are based on songs they have already recognized. The aim of this variant of Hooked is to test whether an adaptive version of the game can be used as a tool to guide players towards the music they are most familiar with. This would make it a valuable instrument in helping music listeners who suffer from

memory loss reconnect with their preferred music even if they have forgotten titles and names of artists.

9.4 THE FUTURE OF AUDIO CORPUS ANALYSIS

In this thesis, we have tried to make a useful contribution to the available methods and technologies for audio description and audio corpus analysis. We have strived to make these methods and technologies transparent and flexible. Along the way, we have gained new insights into choruses, catchiness and hooks.

By presenting concrete applications of the proposed technologies, following the proposed methods, we believe we have shown that rigorous audio corpus analysis is possible and that, even though there is more work left to do, the technologies to engage with it are available.

In pursuing a methodology that is transparent about the origin and limitations of data and algorithms, and technologies that leverage the context and cognition of listening, we make it easier to research music from a critical angle: as a product of the mind, with a strong cultural dimension. We believe this makes our contributions important steps towards the integration of audio analysis into the new empirical method, and, ultimately, musicology.

Hopefully, our efforts can be a stepping stone and an inspiration for future empirical, audio-based research. We encourage researchers to take up this approach, learn from our experiences, and use the wealth of audio available today to discover more about music and our intriguing relation to it as listeners.

A

HOOKED! AUDIO PCA LOADINGS

Feature	Component											
	1	2	3	4	5	6	7	8	9	10	11	12
MIB | Song	0.31	−0.10	0.12	0.08	0.05	0.66	0.05	0.08	0.23	0.08	−0.01	0.14
HI | Song	−0.25	−0.08	0.12	0.06	0.11	0.55	0.12	0.35	−0.06	0.04	0.01	−0.02
MIB | Corpus	0.15	−0.03	−0.02	0.13	0.00	0.77	−0.06	0.00	0.08	−0.02	−0.01	0.05
HI | Corpus	−0.28	−0.09	−0.05	−0.01	0.10	0.55	0.11	0.42	−0.15	−0.02	0.08	−0.05
HIC | Song	0.04	0.13	0.22	0.04	0.00	0.13	−0.04	0.58	−0.03	0.06	−0.02	−0.03
HIC | Corpus	−0.23	0.11	0.04	0.32	0.08	0.15	−0.07	0.66	0.03	−0.06	0.07	0.00
HIC Entropy	0.88	0.06	0.03	−0.16	0.02	0.07	−0.02	−0.23	−0.12	0.02	0.00	−0.10
MIB Entropy	0.83	−0.15	0.00	−0.19	0.04	0.04	0.08	0.26	0.26	0.03	−0.02	0.20
HI Entropy	0.85	−0.06	0.02	−0.20	0.01	−0.01	0.04	0.15	0.12	0.02	−0.02	0.16
HIC Song Information	0.84	0.17	0.06	0.09	0.11	0.13	−0.02	−0.16	−0.28	−0.04	0.10	−0.13
MIB Song Information	0.79	−0.21	−0.03	0.01	0.07	0.05	0.13	0.25	0.29	0.07	−0.02	0.21
HI Song Information	0.90	0.18	0.01	0.11	0.07	−0.07	0.00	−0.17	−0.03	−0.02	0.00	−0.03
HIC Corpus Information	0.86	0.16	0.06	0.01	0.10	0.11	−0.02	−0.20	−0.27	−0.02	0.09	−0.13
MIB Corpus Information	0.79	−0.19	−0.01	−0.03	0.07	0.02	0.14	0.26	0.31	0.07	−0.02	0.21
HI Corpus Information	0.90	0.15	0.02	−0.01	0.03	−0.12	−0.01	−0.24	−0.03	0.00	−0.02	−0.03
HIB Entropy | Song	0.03	0.11	0.42	0.08	0.03	0.00	−0.08	0.15	0.08	0.19	0.01	−0.06
MIB Entropy | Song	0.01	−0.01	0.07	0.10	0.03	−0.01	0.03	0.02	−0.01	0.82	0.00	0.05
HI Entropy | Song	0.03	0.02	0.11	0.12	0.06	0.04	−0.02	−0.01	0.02	0.81	−0.01	0.02
HIB Entropy | Corpus	−0.13	0.08	0.08	0.68	0.08	0.15	−0.06	0.26	−0.03	−0.10	0.07	−0.02
MIB Entropy | Corpus	−0.04	−0.09	−0.01	0.80	0.01	0.06	0.14	−0.01	0.05	0.16	0.00	0.07
HI Entropy | Corpus	−0.03	−0.07	−0.02	0.84	0.04	0.04	0.06	0.04	0.05	0.19	−0.02	0.04
Loudness	−0.04	0.92	0.07	−0.06	−0.05	−0.05	−0.07	0.06	−0.04	0.02	−0.07	0.04
Roughness	0.14	0.78	0.14	0.01	0.15	0.09	0.31	0.06	−0.08	0.07	0.06	0.01
Melodic Pitch Height	0.13	0.66	−0.05	−0.03	0.09	−0.24	−0.16	0.09	0.22	−0.06	−0.06	0.00
MFCC Variance	0.13	−0.51	−0.05	0.08	−0.26	0.10	0.05	−0.02	0.48	0.02	−0.22	−0.10
Loudness | Song	−0.03	−0.05	0.67	−0.01	0.06	0.01	0.07	−0.04	0.10	0.03	0.11	−0.03
Roughness | Song	0.04	0.10	0.67	−0.03	−0.01	0.02	0.11	0.08	−0.05	−0.02	−0.04	−0.05
Mel. Pitch Height | Song	−0.01	0.02	0.46	0.03	0.13	0.14	−0.12	−0.15	0.29	0.07	0.16	0.03
MFCC Mean | Song	0.07	0.07	0.61	−0.04	0.21	0.12	0.10	0.10	−0.07	0.16	0.11	0.11
MFCC Variance | Song	0.00	−0.04	0.54	0.03	0.01	−0.06	0.10	0.08	−0.10	−0.09	−0.06	0.17
Loudness | Corpus	0.04	−0.23	0.06	0.07	0.12	0.08	0.76	−0.05	0.22	0.02	0.10	−0.05
Roughness | Corpus	0.12	0.34	0.15	0.03	0.00	0.01	0.71	−0.07	0.05	0.04	0.03	−0.07
Mel. Pitch Height | Corpus	0.00	0.04	0.06	0.06	0.25	0.06	0.14	−0.01	0.60	0.02	0.14	−0.09
MFCC Mean | Corpus	0.21	0.13	0.12	0.07	0.51	0.03	0.31	0.20	−0.18	0.05	0.14	0.08
MFCC Variance | Corpus	−0.09	−0.09	0.08	0.08	0.25	−0.02	0.40	0.05	−0.13	−0.13	−0.12	0.21
Sharpness	0.23	0.11	0.03	0.08	0.72	0.04	0.29	0.13	0.08	−0.01	0.10	0.05
Sharpness | Song	−0.02	−0.07	0.24	−0.04	0.50	0.06	−0.14	−0.07	0.04	0.15	−0.08	−0.04
Sharpness | Corpus	0.08	0.10	0.03	0.06	0.75	0.03	0.03	−0.02	0.14	−0.01	−0.10	−0.01
Loudness SD	0.10	0.38	0.09	0.06	−0.06	0.06	0.22	0.02	0.40	0.03	−0.61	−0.03
Loudness SD | Song	0.04	0.02	0.22	0.02	−0.05	0.00	−0.05	0.03	0.14	0.01	0.60	0.03
Loudness SD | Corpus	0.03	0.05	−0.02	0.05	−0.03	0.04	0.19	0.02	0.04	0.00	0.78	0.02
Mel. Pitch SD	0.21	−0.10	−0.02	−0.05	0.04	−0.19	0.21	0.18	−0.27	0.12	−0.07	−0.28
Mel. Pitch SD | Song	0.01	0.04	0.11	0.01	0.04	0.13	0.00	−0.15	0.01	0.14	0.07	0.69
Mel. Pitch SD | Corpus	0.13	0.03	−0.02	0.06	−0.01	−0.02	0.01	0.11	−0.08	−0.04	0.00	0.74
R^2	0.16	0.06	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.03

Note. MIB = Melodic Interval Bigram; HI = Harmonization Interval; HIC = Harmony Interval Co-occurrence. Loadings > .40 are in boldface. Collectively, these components explain 64 % of the variance in the underlying data. We interpret and name them as follows: (1) Melodic/Harmonic Entropy, (2) Timbral Intensity, (3) Timbral Recurrence, (4) Melodic/Harmonic Entropy Conventionality, (5) Sharpness Conventionality, (6) Melodic Conventionality, (7) Timbral Conventionality, (8) Harmonic Conventionality, (9) Vocal Prominence, (10) Melodic Entropy Recurrence, (11) Dynamic Range Conventionality, and (12) Melodic Range Conventionality.

Table 8.: Loadings after varimax rotation for principal component analysis of corpus-based audio features.

NEDERLANDSE SAMENVATTING

De relatie tussen de informatica en de menswetenschappen werd de laatste tien jaar gekenmerkt door de snelle groei van de 'digital humanities'. Dat is een interdisciplinair onderzoeksveld waarin de onderzoeksmethoden van de twee disciplines samenkomen. De opmars van deze digital humanities wordt meestal toegeschreven aan de beschikbaarheid van ongekende hoeveelheden data en de juiste instrumenten om deze te analyseren.

In de taalwetenschappen bijvoorbeeld, is het nu eenvoudiger dan ooit om een bepaalde hypothese te testen op steeds grotere corpora. De onderzoekers die hier gebruik van maken hebben dit te danken aan het werk van computerpioniers die zich al vroeg in het computertijdperk toewijdden aan het digitaliseren van data en het ontwikkelen van standaarden en infrastructuur. Ook in de musicologie begon het digitale werk al in de jaren zestig en zeventig, met onder andere projecten waarin belangrijke werken uit de Renaissance werden gedigitaliseerd.

Aan het einde van de jaren negentig geraakte het digitale en computationele muziekonderzoek in een stroomversnelling door de opkomst van het internet en nieuwe technieken voor digitale signaalverwerking. Het huidige *music information retrieval* vakgebied (MIR) is hiervan het resultaat. De drijfveer voor dit soort onderzoek is echter vaak praktischer van aard: MIR streeft er vooral naar muziek beter doorzoekbaar te maken, met toepassingen als het classificeren van muziek in een collectie en het aanraden van nieuwe muziek aan gebruikers van een streamingdienst.

Maar er is ook aandacht voor musicologische vragen. Onderzoekers met uiteenlopende achtergronden zijn MIR technieken blijven gebruiken om op steeds grotere schaal op zoek te gaan naar nieuwe inzichten. Voorbeelden van dit soort onderzoek, op basis van 'corpusanalyse', zijn te vinden in het werk van cognitiewetenschappers

David Huron [80] en Marc Leman (bv. [104]). In het onderzoek naar populaire muziek zijn er de voor dit proefschrift belangrijke studies van Serrà et al., [175] en Mauch et al. [122], waarin gekeken wordt naar de evolutie van popmuziek over de laatste vijftig jaar. In een recent onderzoek van Savage et al., ten slotte, wordt in een wereldwijd corpus gezocht naar universele kenmerken van muziek [171].

Wat opvalt in deze studies—een overzicht wordt gegeven in hoofdstuk 2—is dat in veruit het grootste deel ervan geen gebruik wordt gemaakt van audiodata, maar van 'symbolische data': partituren, akkoorden of handmatig toegekende labels. Dit terwijl het meeste onderzoek in MIR net gedaan wordt met behulp van audiodata, het formaat waarin er eenvoudigweg veel meer muziek beschikbaar is. Kortom, ondanks de beschikbaarheid van muziekopnames en de hulpmiddelen voor het analyseren ervan, is er maar weinig onderzoek verricht naar de corpusanalyse van audiodata.

Daar wilde ik met mijn onderzoek iets aan doen. Hoofdstuk 1 gaat dieper in op de motivatie hiervoor, en bespreekt de wetenschappelijke context waarin corpusanalyse kan worden gesitueerd. De rest van het proefschrift presenteert een aantal concrete bijdragen aan aan het wetenschappelijke muziekonderzoek op basis van *audiocorpusanalyse*. Ik concentreer me hierbij op drie themas: audiobeschrijving, methodieken voor corpusanalyse, en het toepassen van deze beschrijving- en analysetechnieken in een onderzoek naar 'hooks'. De eerste twee themas zijn verweven in deel i en ii van deze thesis, het derde thema wordt behandeld in deel iii. In de rest van deze samenvatting licht ik mijn werk rond elk van deze thema's kort toe.

Audiobeschrijving

Hoofdstuk 2 van dit proefschrift gaat over manieren om de inhoud van een audiofragment in getallen te vatten. Dat is namelijk nodig om op grote schaal audio te analyseren: de data moeten vertaald worden naar een voorstellingswijze waarmee de computer kan gaan rekenen. Gelukkig zijn er in het MIR vakgebied honderden methoden ontwikkeld om verschillende soorten eigenschappen van audiodata te meten.

Deze gaan van, op het 'laagste' niveau, concrete eigenschappen van het signaal (bv. de totale energie of het frequentiespectrum) tot muziektheoretische eigenschappen zoals de complexiteit van eventuele ritmische patronen of akkoorden. De eerste helft van het hoofdstuk geeft een overzicht van deze meetbare eigenschappen of 'features', met een nadruk op het beschrijven van melodie, harmonie en timbre.

In hoofdstukken 5 en 6 wordt verder ingegaan op dit onderwerp. Hier wordt een verzameling nieuwe features voorgesteld, verder genaamd 'audio bigrams'. Ze zijn geinspireerd door sommige van de features die gebruikt worden in het beschrijven van symbolische data, maar met een aantal belangrijke verschillen. Veel symbolische features zijn namelijk gebaseerd op technieken uit de tekstanalyse, waarin documenten gemakkelijk opgedeeld kunnen worden in woorden en zinnen. Muziekbronnen in symbolische formaten kunnen vaak op een vergelijkbare manier opgebroken worden in noten en akkoorden of motieven, die dan geteld worden, of geïndexeerd.

Audiodata laten dit niet gemakkelijk toe: het is vaak moeilijk te zeggen waar het ene geluid eindigt en het andere begint, of waar er een maatstreep valt. De typische oplossing hiervoor is om elk audiofragment dan maar op te delen in miniscule 'frames' met een vaste lengte en op een vaste afstand van elkaar. De manieren om de informatie uit elk van die frames daarna te combineren lopen echter uiteen. In veel gevallen wordt er niet eens gekeken naar de volgorde van de frames. Daarmee wordt natuurlijk veel informatie weggegooid over de volgorde waarin dingen gebeuren—iets wat vanzelfsprekend belangrijk is in de muziek.

Het alternatief dat in dit proefschrift wordt verdedigd is een breed toepasbaar model genaamd 'audio bigrams'. Features gebaseerd op het audio bigram model beschrijven hoe vaak twee gegeven observaties voorkomen in frames die dicht bij elkaar liggen in het audiofragment. Die observaties kunnen bijvoorbeeld akkoorden zijn (F groot en C groot), of 'trappen' in de melodie (IV en I). De formele definitie van het audio bigram model laat niet alleen toe weer te geven hoe vaak twee observaties samen voorkomen, maar ook in welke volgorde. Dat maakt de features erg bruikbaar voor corpusanalyse.

In hoofdstuk 5 wordt bij wijze van evaluatie nagegaan in welke mate audio bigram features nuttig zijn om volledige liedjes te beschrijven. Er wordt aangetoond dat op basis van slechts enkele eenvoudige features per liedje, 'covers' kunnen gevonden worden in een kleine dataset—een vorm van 'fingerprinting'.

In hoofdstuk 6 wordt vervolgens geargumenteerd dat het berekenen van audio bigram features in veel gevallen volledig 'gevectoriseerd' kan worden. Dat wil onder andere zeggen dat ze erg efficient te berekenen zijn. Daarna wordt de gelijkenis met fingerprinting—het efficiënt identificeren van een geluidsfragment op basis van features—verder doorgetrokken. Verschillende technieken voor het herkennen van korte audiofragmenten, covers of samples kunnen herleid worden tot iets wat sterk op audio bigram features lijkt. Dit suggereert dat het audio bigram model een nieuwe manier van fingerprinting mogelijk maakt waarbij, met de juiste dataset, de benodigde features volledig vanzelf door het systeem kunnen ontdekt worden.

In hoofdstuk 8 worden opnieuw een aantal nieuwe technieken voor audiobeschrijving voorgesteld, onder de verzamelnaam 'second-order features'. Deze features quantificeren hoe typisch of atypisch een bepaalde combinatie aan featurewaarden is, en kunnen dus gebruikt worden om een typische melodie van een atypische melodie, of een typisch timbre van een atypisch timbre te onderscheiden in een muziekfragment.

Corpusanalyse

In dit proefschrift wordt op drie plaatsen een bijdrage geleverd aan corpusanalyse als onderzoeksmethode. Aan het einde van deel i (hoofdstuk 3) wordt eerst een overzicht gegeven van de verschillende soorten corpusanalyse die in de voorbije eeuw naar voren zijn geschoven, te beginnen met het werk van Alan Lomax, een musicoloog die tussen 1930 en 1960 duizenden veldopnamen maakte van volkszangers en muzikanten in de Verenigde Staten, Europa, de Caraïbische eilanden en verschillende andere plaatsen. Later, in de jaren zestig en zeventig, gebruikte hij statistische analyse om de eigenschappen van

zijn opnames te correleren aan andere soorten etnografische informatie [112].

Met de opkomst van de computer werd er steeds meer van dit soort 'data-intensief' onderzoek verricht, bijvoorbeeld naar het gebruik van akkoorden in rock en popmuziek [27, 41]. Ook in de muziekcognitie, een tak van de cognitiewetenschappen, werd aan corpusanalyse gedaan, bv. door David Huron, die onderzoek doet naar muzikale verwachtingspatronen. Cognitiewetenschapper Marc Leman deed naar de relatie tussen muziek en wandelsnelheid, en muziekpsycholoog Daniel Müllensiefen deed onderzoek naar muziek en het geheugen [80, 104, 134]. Op het onderzoek van Leman na zijn al deze studies gebaseerd op symbolische data of handmatig toegekende labels.

Het overzicht van de verschillende corpusanalysemethoden wordt daarom gevolgd door een case study rond twee artikels, van Serrà et al. en Mauch et al., waarin de evolutie van populaire muziek wordt geanalyseerd [122, 175]. De conclusies van de twee artikels spreken elkaar enigszins tegen: Serrà argumenteert dat muziek steeds meer 'homogeen' is geworden in zowel harmonie als timbre, maar Mauch vindt hier geen aanwijzingen voor.

In het laatste deel van hoofdstuk 3 worden de conclusies van het literatuuroverzicht en de casestudie samengebracht in negen aanbevelingen voor goed corpus-gebaseerd onderzoek. De lijst omvat aanbevelingen rond de keuze van hypotheses, data, features en statistische modellen. De meest belangrijke hiervan raken aan gekende valkuilen: leg eventuele hypothesen vast voor de data geanalyseerd worden, zorg voor genoeg representatieve data en houd rekening met mogelijk misleidende correlaties (*spurious correlations*). Overige aanbevelingen gaan onder andere over het gebruik van features. Zo luidt één van de conclusies dat verschillende problemen voorkomen kunnen worden als het aantal features beperkt is.

In hoofdstuk 4 wordt een eerste corpusanalyse gepresenteerd. De onderzoeksvraag was: hoe verschillen refreinen van andere delen van een popnummer? Als features werd er gekozen voor een kleine verzameling veelgebruikte MIR features waar een eenvoudige interpretatie voor bestaat. Twee datasets werden gebruikt: een kleine dataset met

oude Nederlandstalige nummers, uitgebracht tussen 1905 en 1955, en een dataset met 7762 fragmenten van 649 nummers uit Ashley Burgoyne's Billboard dataset, uitgebracht tussen 1957 en 1992. Het meest complexe deel van deze analyse was het gekozen statistisch model, een zogenaamd *probabilistic graphical model* (PGM). De resultaten laten zien dat de gemiddelde toonhoogte in refreinen hoger is dan in strofen, en dat refreinen een meer prominente melodie bevatten.

Hooks

Het derde en laatste deel van dit proefschrift gaat over de muzikale eigenschappen van 'hooks'. De term hook komt uit de songwritingliteratuur en verwijst naar het meest 'catchy', gemakkelijk te onthouden deel van een popnummer. Dat kan het refrein zijn, een deel van het refrein, of een heel ander element in het nummer, en de opvattingen over wat een hook tot hook maakt lopen in de literatuur nogal uiteen. Dat alles maakt hooks een ideaal onderwerp voor corpusanalyse. En aangezien popmuziek zich ook nog eens moeilijk laat omschrijven in een symbolisch dataformaat (meer hierover in hoofdstuk 1), is met name het analyseren van een audiocorpus hier een goede uitweg.

Maar voor dit mogelijk is, zijn er eerst data nodig, en data verzamelen over hooks—welke fragmenten zijn hooks en welke niet—is een aanzienlijke uitdaging. Samen met collega's in het COGITCH project werd besloten hooks te definiëren als het deel van een nummer dat het meest herkenbaar is. 'Herkenbaarheid' werd gedefinieerd aan de hand van een geheugenmodel waarin 'herkenningstijd' centraal staat: de tijd die de luisteraar nodig heeft om met zekerheid te kunnen zeggen, 'dit nummer ken ik'! Verder werd ervoor gekozen schattingen van deze herkenningstijd te verzamelen in een grootschalig internetexperiment, dat de vorm kreeg van een game. Al deze keuzes worden verder toegelicht in hoofdstuk 7.

Gedurende de laatste drie jaar zijn er verschillende versies van de game tot stand gekomen. De eerste versie, *Hooked!* was gericht op

een Nederlands publiek.[1] Deelnemers krijgen verschillende muziekfragementen te horen, telkens met de vraag: 'Ken je dit nummer'? (Zie ook de screenshots in figuur 28.) Bij elk ja-antwoord volgt er een tweede vraag die test of de deelnemer weet hoe het nummer verder gaat. Om en bij de tweeduizend personen speelden *Hooked!*, en de centrale vraag in het spel werd in totaal 167.000 keer beantwoord.

Een vernieuwde versie van het spel, *Hooked on Music*[2], werd gemaakt in samenwerking met Manchester Science Festival.[3] Deze versie was gericht op een internationaal publiek. Na een persbericht over de eerste resultaten van dit experiment vonden meer dan 160.000 deelnemers hun weg naar het spel, samen goed voor meer dan 3 miljoen antwoorden op de vraag, 'Do you know this song?'.

Hoofdstuk 8 presenteert de corpusanalyse van de data uit *Hooked!*. Dezelfde features werden gebruikt als in de analyse van de refreinen in hoofdstuk 4, aangevuld met een verzameling features gebaseerd op het audio bigram model uit hoofdstuk 5 en een aantal second-order features (zie boven). Om relaties tussen deze features en de data uit de game (de herkenbaarheid van de fragmenten) bloot te leggen, werd gebruik gemaakt van *principal components analysis* en *linear mixed effects regression*. De resultaten laten zien dat de delen van een nummer waarin een stem voorkomt en de delen die een representatief timbre hebben voor de rest van het nummer het meest herkenbaar zijn. Hooks hebben verder ook een meer typische melodie, harmonie en timbre.

Tenslotte

Met dit proefschrift heb ik een poging gedaan bij te dragen aan de beschikbare methodes en technologieën voor het beschrijven en analyseren van audiocorpora. Door deze methodes en technologieën toe te passen in concrete experimenten, met concrete resultaten, geloof

1 Kijk op `http://hooked.humanities.uva.nl/` voor meer informatie en op `http://www.hookedgame.nl` voor de game zelf.

2 `http://www.hookedonmusic.org.uk/`

3 `http://www.manchestersciencefestival.com/`

ik aangetoond te hebben dat corpusanalyse van audiodata kan leiden tot robuuste, bruikbare kennis. Ik hoop dan ook dat de bovenstaande resultaten het begin kunnen zijn van een lange reeks nieuwe inzichten in de eigenschappen van hooks, herkenbaarheid en ons geheugen voor muziek.

CURRICULUM VITAE

Jan Van Balen (October 26, 1988) grew up in Leuven, Belgium, where he studied music at the city conservatory, and science and mathematics in high school. In 2002–2003 he spent 9th grade at Palisades High School in Los Angeles.

Between 2006 and 2009, Jan completed a bachelor in Physics at the Katholieke Universiteit Leuven. Afterwards, he started a preparatory master in Mathematical Engineering at the same university.

In 2010, he moved to Barcelona, Spain, to study a master in Sound and Music Computing at Universitat Pompeu Fabra, with subjects in music information retrieval, music perception and cognition, and music production. He finished the master in 2011 with a thesis project on the automatic recognition of samples in musical audio, under supervision of dr. Joan Serrà.

A continued interest in the analysis of reuse in music led him to his PhD position at Utrecht University, in the COGITCH project, where his research has focused on audio corpus analysis and popular music. Other research topics he worked on are hooks and music memory, audio content identification, and stability and variation in folk and popular music.

BIBLIOGRAPHY

[1] Hans Abbing. Een speelse canon voor de popmuziek. In Huub B M Wijfjes, editor, *De muziek zegt alles. De Top 2000 onder professoren*. Uitgeverij L J Veen, Amsterdam, 2011.

[2] Charles R. Adams. Melodic Contour Typology. *Ethnomusicology*, 20(2):179–215, 1976.

[3] Anna Aljanaki, Dimitrios Bountouridis, John Ashley Burgoyne, Frans Wiering, Jan Van Balen, and Henkjan Honing. Designing Games with a Purpose for Music Research: Two Case Studies. In *The Games And Learning Alliance Conference (GALA)*, Paris, France, 2013.

[4] Anna Aljanaki and Frans Wiering. Computational Modeling Of Induced Emotion Using Gems. In *Proc. 15th International Society for Music Information Retrieval Conference*, pages 373–378, 2014.

[5] Jean-Julien Aucouturier. Sounds Like Teen Spirit: Computational Insights into the Grounding of Everyday Musical Terms. *Language, Evolution and the Brain*, pages 35–64, 2009.

[6] Jean-Julien Aucouturier and Emmanuel Bigand. Seven Problems That Keep MIR From Attracting the Interest of Cognition and Neuroscience. *Journal of Intelligent Information Systems*, 41(3):483–497, jul 2013.

[7] Jean-Julien Aucouturier and François Pachet. Representing Musical Genre : A State of the Art Representing Musical Genre : A State of the Art. *Journal of New Music Research*, 32(1):83–93, 2003.

[8] Mathieu Barthet, Mark D Plumbley, Alexander Kachkaev, Jason Dykes, Daniel Wolff, and Tillman Weyde. Big Chord Data Ex-

traction and Mining. In *Proc. 9th Conference on Interdisciplinary Musicology*, 2014.

[9] Mark A. Bartsch and Gregory H. Wakefield. To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 15–18. IEEE, 2001.

[10] Mark A. Bartsch and Gregory H. Wakefield. Audio Thumbnailing of Popular Music Using Chroma-Based Representations. *IEEE Transactions on Multimedia*, 7(1):96–104, 2005.

[11] James Bergstra, Norman Casagrande, Dumitru Erhan, Douglas Eck, and Balázs Kégl. Aggregate features and ADABOOST for music classification. *Machine Learning*, 65(2-3):473–484, 2006.

[12] Thierry Bertin-Mahieux and Daniel P. W. Ellis. Large-scale Cover Song Recognition Using Hashed Chroma Landmarks. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 10–13, 2011.

[13] Thierry Bertin-Mahieux and Daniel P. W. Ellis. Large-Scale Cover Song Recognition Using The 2d Fourier Transform Magnitude. In *Proc. 13th International Society for Music Information Retrieval Conference*, pages 2–7, 2012.

[14] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In *Proc. 12th International Society for Music Information Retrieval Conference*, 2011.

[15] Jacob Bien and Robert J. Tibshirani. Sparse Estimation of a Covariance Matrix. *Biometrika*, 98(4):807–820, 2011.

[16] Christopher M. Bishop. Graphical Models. In *Pattern Recognition and Machine Learning*, chapter 8. Springer, 2006.

[17] Dimitrios Bountouridis and Jan Van Balen. The Cover Song Variation Dataset. In *Proc. 4th International Workshop on Folk Music Analysis*, pages 5–7, Istanbul, Turkey, 2014.

[18] Dimitrios Bountouridis and Jan Van Balen. Towards Capturing Melodic Stability. In *Proc. 9th Conference on Interdisciplinary Musicology*, Berlin, Germany, 2014.

[19] Judith C. Brown. Calculation of a Constant Q spectral Transform. *The Journal of the Acoustical Society of America*, 89(1):425, 1991.

[20] Scott Brown and Andrew Heathcote. The Simplest Complete Model of Choice Response Time: Linear Ballistic Accumulation. *Cognitive Psychology*, 57(3):153–178, 2008.

[21] John Ashley Burgoyne. *Stochastic Processes & Database-driven Musicology*. PhD thesis, McGill University, 2011.

[22] John Ashley Burgoyne. Music Information Retrieval. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A New Companion to Digital Humanities*, pages 213—-228. Wiley-Blackwell, 2016.

[23] John Ashley Burgoyne, Dimitrios Bountouridis, Jan Van Balen, and Henkjan Honing. Hooked: a Game for Discovering What Makes Music Catchy. In *Proc. 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[24] John Ashley Burgoyne, Abe D. Hofman, Han L.J. van der Maas, and Henkjan Honing. Adaptive Music Recognition Games for Dementia Therapy. In *Proceedings of the European Society for the Cognitive Sciences of Music*, Manchester, England, 2015.

[25] John Ashley Burgoyne, Jan Van Balen, Dimitrios Bountouridis, Themistoklis Karavellas, Frans Wiering, Remco C. Veltkamp, and Henkjan Honing. The Contours of Catchiness, or Where to Look for a Hook. In *International Conference on Music Perception and Cognition*, Seoul, South Korea, 2014.

[26] John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. An Expert Ground-truth Set For Audio Chord Recognition and Music Analysis. In *Proc. 12th International Society Music Information Retrieval Conference*, 2011.

[27] John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. Compositional Data Analysis of Harmonic Structures in Popular Music. *Mathematics and Computation in Music*, pages 52–63, 2013.

[28] Gary Burns. A Typology of 'Hooks' in Popular Records. *Popular music*, 6(1):1–20, 1987.

[29] Marcelo Caetano and Frans Wiering. Theoretical Framework of a Computational Model of Auditory Memory for Music Emotion Recognition. *Proc. 15th International Society for Music Information Retrieval Conference*, pages 331–336, 2014.

[30] Murray Campbell. Timbre (i). *Grove Music Online. Oxford Music Online.*, 2016.

[31] Pedro Cano, Eloi Battle, Ton Kalker, and Jaap Haitsma. A Review of Algorithms for Audio Fingerprinting. *IEEE Workshop on Multimedia Signal Processing.*, pages 169–173, 2002.

[32] Wei Chai. *Automated Analysis of Musical Structure*. PhD thesis, Massachusetts Institute of Technology, Boston, Massachusetts, 2005.

[33] Elaine Chew. *Towards a Mathematical Model of Tonality*. PhD thesis, Massachusetts Institute of Technology, 2000.

[34] Michael Clausen and Meinard Müller. Transposition-Invariant Self-Similarity Matrices. In *Proc. 8th International Society for Music Information Retrieval Conference*, 2007.

[35] Adam Coates, Ann Arbor, and Andrew Y Ng. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. *Aistats 2011*, pages 215–223, 2011.

[36] Nicholas Cook. *Music: A Very Short Introduction*. Oxford Paperbacks, Oxford, 1998.

[37] Nicholas Cook and Eric Clarke. What is Empirical Musicology? In Eric Clarke and Nicholas Cook, editors, *Empirical Musicology: Aims, Methods, Prospects*, chapter 1, pages 3—-14. Oxford University Press, Oxford, New York, 2004.

[38] Roger B. Dannenberg and Masataka Goto. Music Structure Analysis from Acoustic Signals. In David Havelock, Sonoko Kuwano, and Michael Vorländer, editors, *Handbook of Signal Processing in Acoustics*. Springer New York, 2008.

[39] Roger B. Dannenberg and Ning Hu. Pattern Discovery Techniques for Music Audio. *Journal of New Music Research*, 32(2):153–163, jun 2003.

[40] Alain de Cheveigne and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917, 2002.

[41] Trevor de Clercq and David Temperley. A corpus analysis of rock harmony. *Popular Music*, 30(01):47–70, jan 2011.

[42] Emmanuel Deruty and Damien Tardieu. About dynamic processing in mainstream music. *Journal of the Audio Engineering Society*, 62(1-2):42–56, 2014.

[43] Sander Dieleman, Philémon Brakel, and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. *Proc. 12th International Society for Music Information Retrieval Conference*, pages 669–674, 2011.

[44] Sander Dieleman and Benjamin Schrauwen. Multiscale Approaches to Musical Audio Feature Learning. In *Proceedings of the 14th international society for music information retrieval conference*, 2013.

[45] Christian Dittmar, Kay F. Hildebrand, Daniel Gaertner, Manuel Winges, Florian Müller, and Patrick Aichroth. Audio Forensics Meets Music Information Retrieval - A Toolbox for Inspection

of Music Plagiarism. In *European Signal Processing Conference (EUSIPCO)*, pages 1249–1253, 2012.

[46] William Fishburn Donkin. *Acoustics*. Macmillan and Co., London, 1870.

[47] J. Stephen Downie. The music information retrieval evaluation exchange (20052007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247—–255, 2008.

[48] Daniel P.W. Ellis and Courtenay V Cotton. The 2007 Labrosa Cover Song Detection System. In *MIREX 2007*, 2007.

[49] Daniel P.W. Ellis, C.V. Cotton, and M.I. Mandel. Cross-correlation of beat-synchronous representations for music similarity. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 57–60. IEEE, 2008.

[50] A Eronen and F Tampere. Chorus detection with combined use of MFCC and chroma features and image processing filters. In *Proc Int Conf Digital Audio Effects (DAFx)*, pages 1–8, 2007.

[51] Sébastien Fenet, Richard Gaël, and Yves Grenier. A Scalable Audio Fingerprint Method with Robustness to Pitch-shifting. In *Proc. 12th International Conference on Music Information Retrieval*, 2011.

[52] William Tecumseh Fitch. Four principles of bio-musicology. *Philosophical Transactions of the Royal Society B*, 10(3):197–202, 2015.

[53] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[54] Takuya Fujishima. Realtime chord recognition of musical sound: A system using common lisp music. In *Proc. International Computer Music Conference*, pages 464–467, 1999.

[55] Emilia Gomez. *Tonal Description of Musical Audio Signals*. PhD thesis, Universitat Pompeu Fabra, 2006.

[56] Emilia Gomez and Perfecto Herrera. Comparative Analysis of Music Recordings from Western and Non-Western Traditions by Automatic Tonal Feature Extraction and Data Mining. *Empirical Musicology Review*, 3(3):68, 2008.

[57] Masataka Goto. A chorus-section detecting method for musical audio signals. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 437—-440, 2003.

[58] Masataka Goto. A Chorus Section Detection Method for Musical Audio Signals and Its Application to a Music Listening Station. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1783–1794, 2006.

[59] Peter Grosche, Meinard Müller, and Joan Serrà. Audio Content-based Music Retrieval. In Meinhard Müller, Masataka Goto, and Markus Schedl, editors, *Multimodal Music Processing*. Dagstuhl Publishing, 2012.

[60] Enric Guaus. *Audio content processing for automatic music genre classification: descriptors, databases, and classifiers.* PhD thesis, Universitat Pompeu Fabra, 2009.

[61] W. Bas De Haas and Frans Wiering. Hooked on Music Information Retrieval. *Empirical Musicology Review*, 5(4):176–185, 2010.

[62] Jaap Haitsma, Ton Kalker, and Job Oostveen. Robust audio hashing for content identification. In *International Workshop on Content-Based Multimedia Indexing*, pages 117–124. Citeseer, 2001.

[63] Philippe Hamel and Douglas Eck. Learning Features from Music Audio with Deep Belief Networks. In *Proc. 11th International Society for Music Information Retrieval Conference*, pages 339–344, 2010.

[64] Philippe Hamel, Simon Lemieux, Yoshua Bengio, and Douglas Eck. Temporal Pooling and Multiscale Learning for Automatic Annotation and Ranking of Music Audio. *Proc. 12th International Society for Music Information Retrieval Conference*, page 876, 2011.

[65] Christopher Harte, Mark Sandler, and Martin Gasser. Detecting harmonic change in musical audio. *Proceedings of the 1st ACM workshop on Audio and music computing multimedia AMCMM 06*, C(06):21, 2006.

[66] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[67] Wolfgang Hess. *Pitch Determination of Speech Signals*. Springer Science & Business Media, 1983.

[68] Henkjan Honing. The comeback of systematic musicology: new empiricism and the cognitive revolution. *Tijdschrift voor Muziektheorie*, 9(3):241, 2004.

[69] Henkjan Honing. Lure(d) into listening : The potential of cognition-based music information retrieval. *Empirical Musicology Review*, 5(4):146–151, 2010.

[70] Henkjan Honing. *The Illiterate Listener: On Music Cognition, Musicality and Methodology*. Vossiuspers UvA, Amsterdam, 2011.

[71] Henkjan Honing. *Musical cognition: a science of listening*. Transaction Publishers, 2014.

[72] John L Horn. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*, 30(2):179–185, 1965.

[73] Yifan Hu, Chris Volinsky, and Yehuda Koren. Collaborative filtering for implicit feedback datasets. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 263–272, 2008.

[74] Eric J. Humphrey, Aron P. Glennon, and Juan Pablo Bello. Nonlinear semantic embedding for organizing large instrument sam-

ple libraries. *Proc. 10th International Conference on Machine Learning and Applications, ICMLA 2011*, 2:142–147, 2011.

[75] Eric J. Humphrey, Oriol Nieto, and Juan Pablo Bello. Data Driven and Discriminative Projections for Large-Scale Cover Song Identification. In *Proc. 14th International Society for Music Information Retrieval Conference*, 2013.

[76] David Huron. Voice Denumerability in Polyphonic Music of Homogeneous Timbres. *Music Perception: An Interdisciplinary Journal*, 6(4):361—-382, 1989.

[77] David Huron. The Melodic Arch in Western Folksongs. *Computing in musicology*, 1996.

[78] David Huron. Musical Expectation. In *The 1999 Ernest Bloch Lectures*. Ohio State University, 1999.

[79] David Huron. The New Empiricism: Systematic Musicology in a Postmodern Age. *The 1999 Ernest Bloch Lectures*, 1999.

[80] David Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. Bradford Bks. MIT Press, 2006.

[81] David Huron. On the Virtuous and the Vexations in the Age of Big Data. *Music Perception*, 31(1):4—-9, 2013.

[82] David Huron and Elizabeth Hellmuth Margulis. Musical expectancy and thrills. In Patrik N Juslin and John Sloboda, editors, *Handbook of music and emotion: Theory, research, applications*, pages 575–604. Oxford University Press, 2011.

[83] Charles Inskip and Frans Wiering. In Their Own Words: Using Text Analysis to Identify Musicologists Attitudes Towards Technology. In *Proc. 16th International Society for Music Information Retrieval Conference*, pages 455–461, 2015.

[84] Kelly Joan Jakubowski. *Investigating Temporal and Melodic Aspects of Musical Imagery*. PhD thesis, Goldsmiths, University of London, 2015.

[85] Florian Kaiser and Geoffroy Peeters. A Simple Fusion Method of State and Sequence Segmentation for Music Structure Discovery. *Proc. 14th International Society for Music Information Retrieval Conference*, 2013.

[86] Markus Kalisch and Martin Mächler. Causal Inference using Graphical Models with the R-Package pcalg. *Journal of Statistical Software*, 47(11):1—-26, 2012.

[87] Michael Kassler. *Proving musical theorems I: The middleground of Heinrich Schenker's theory of tonality*. Basser Dept. of Computer Science, School of Physics, University of Sydney, 1975.

[88] Joseph Kerman. *Contemplating music: Challenges to musicology*. Harvard University Press, 1985.

[89] Samuel Kim and Shrikanth Narayanan. Dynamic chroma feature vectors with applications to cover song identification. *IEEE 10th Workshop on Multimedia Signal Processing*, pages 984–987, oct 2008.

[90] Samuel Kim, Erdem Unal, and Shrikanth Narayanan. Music fingerprint extraction for classical music cover song identification. *IEEE Multimedia and Expo*, 2008.

[91] Youngmoo E. Kim, Erik M Schmidt, Raymond Migneco, Brandon G. Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A. Speck, and Douglas Turnbull. Music Emotion Recognition : a State of the Art Review. In *Proc. 11th International Society for Music Information Retrieval Conference*, pages 255–266, 2010.

[92] Jacques Klöters. *Omdat ik zoveel van je hou. Nederlandse chansons en cabaretliederen 1895-1958*. Nijgh en Van Ditmar, Amsterdam, 1991.

[93] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, Massachusetts, 2009.

[94] Hendrik Vincent Koops, Anja Volk, and W Bas De Haas. Corpus-Based Rhythmic Pattern Analysis of Ragtime Syncopation. *Proc. 16th International Society for Music Information Retrieval Conference*, pages 483–489, 2015.

[95] Reinhard Kopiez and Daniel Müllensiefen. Auf der Suche nach den 'Popularitätsfaktoren' in den Song-Melodien des Beatles-Album Revolver, 2013.

[96] Mark D. Korhonen, David A. Clausi, and M. Ed Jernigan. Modeling emotional content of music using system identification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(3):588–599, 2006.

[97] Katerina Kosta, Oscar F. Bandtlow, and Elaine Chew. Practical Implications of Dynamic Markings in the Score: is Piano always Piano? In *Proceedings of Audio Engineering Society 53rd International Conference*, 2014.

[98] Peter Van Kranenburg, Martine De Bruin, Louis P Grijp, Frans Wiering, Peter Van Kranenburg, Martine De Bruin, Louis P Grijp, and Frans Wiering. Meertens Online Reports: The Meertens Tune Collections. Technical report, Meertens Instituut, Amsterdam, 2014.

[99] Charles Kronengold. Accidents, hooks and theory. *Popular Music*, 24(03):381, 2005.

[100] Carol L. Krumhansl. *Cognitive foundations of musical pitch*, volume 17. Oxford University Press New York, 1990.

[101] Carol L. Krumhansl. Plink: "Thin Slices" of Music. *Music Perception*, 27(5):337–354, 2010.

[102] Frank Kurth and Meinard Müller. Efficient Index-Based Audio Matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, feb 2008.

[103] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *Proc. International Conference on Digital Audio Effects*, pages 237–244, 2007.

[104] Marc Leman, Dirk Moelants, Matthias Varewyck, Frederik Styns, Leon van Noorden, and Jean-Pierre Martens. Activating and relaxing music entrains the speed of beat synchronized walking. *PloS One*, 8(7), jan 2013.

[105] Fred Lerdahl. Genesis and Architecture of the GTTM Project. *Music Perception*, 26(3):187–194, 2009.

[106] Armand M. Leroi. Digitizing the Humanities, 2015.

[107] Micheline Lesaffre, Marc Leman, Koen Tanghe, B De Baets, De Baets, and Hans De Meyer. User-Dependent Taxonomy of Musical Features As a Conceptual Framework for Musical Audio-Mining Technology. In *Proceedings of the Stockholm Music and Acoustics Conference*, 2003.

[108] Daniel J. Levitin and Perry R. Cook. Memory for musical tempo: additional evidence that auditory memory is absolute. *Perception & psychophysics*, 58(6):927–935, 1996.

[109] Tao Li and Mitsunori Ogihara. Detecting emotion in music. In *Proc. 4th International Symposium on Music Information Retrieval*, pages 239–240, 2003.

[110] Yipeng Li and David Huron. Melodic modeling: A comparison of scale degree and interval. In *Proceedings of the International Computer Music Journal*, 2006.

[111] Peter H Lindsay and Donald A Norman. *Human information processing: An introduction to psychology*. Academic Press, 1977.

[112] Alan Lomax. *Folk Song Style and Culture*. Transaction Books, 1968.

[113] Alan Lomax and Norman Berkowitz. The Evolutionary Taxonomy of Culture. *Science*, 177(4045):228–239, 1972.

[114] Lie Lu, Muyuan Wang, and Hong-Jiang Zhang. Repeating Pattern Discovery and Structure Analysis from Acoustic Music Data. *ACM Workshop on Multimedia Information Retrieval*, pages 0–7, 2004.

[115] Michael I. Mandel and Daniel P. W. Ellis. Song-level features and support vector machines for music classification. *Proc. 11th International Society for Music Information Retrieval Conference*, pages 594–599, 2005.

[116] Christopher Manning and Hinrich Schütze. Collocations. *Foundations of Statistical Natural Language Processing*, pages 151–189, 1999.

[117] Elizabeth Hellmuth Margulis. Musical Repetition Detection Across Multiple Exposures. *Music Perception*, 29(4):377–385, 2012.

[118] Elizabeth Hellmuth Margulis. *On Repeat: How Music Plays the Mind*. Oxford University Press, 2013.

[119] Elizabeth Hellmuth Margulis. Verbatim repetition and musical engagement. *Psychomusicology: Music, Mind and Brain*, 24(2):157, 2014.

[120] Matthias Mauch and Simon Dixon. Approximate note transcription for the improved identification of difficult chords. In *Proc. 11th International Society for Music Information Retrieval Conference*, pages 135–140, 2010.

[121] Matthias Mauch and Simon Dixon. A Corpus-based Study of Rhythm Patterns. In *Proc. 13th International Society for Music Information Retrieval Conference*, pages 163–168, 2012.

[122] Matthias Mauch, Robert M. MacCallum, Mark Levy, and Armand M. Leroi. The evolution of popular music: USA 1960-2010. *Proceedings of the Royal Society B*, 2015.

[123] Stephen McAdams, Philippe Depalle, and Eric Clarke. Analyzing Musical Sound. In Eric Clarke and Nicholas Cook, editors, *Empirical Musicology: Aims, Methods, Prospects*, chapter 8, pages 157——196. Oxford University Press, Oxford, New York, 2004.

[124] Susan McClary. *Conventional wisdom: The content of musical form*. University of California Press, 2000.

[125] Brian Mcfee and Daniel P.W. Ellis. Analyzing Song Structure with Spectral Clustering. In *Proc. 15th International Society for Music Information Retrieval Conference*, 2014.

[126] Peter Mercer-Taylor. Two-and-a-half centuries in the life of a Hook. *Popular Music & Society*, 23(2):1–15, 1999.

[127] Paul Mermelstein. Distance measures for speech recognition, psychological and instrumental, 1976.

[128] Leonard B. Meyer. *Emotion and Meaning in Music*. Phoenix books. University of Chicago Press, 1961.

[129] Richard Middleton. Form. In J. Shepherd, D. Horn, and D. Laing, editors, *Continuum Encyclopedia of Popular Music of the World Part 1: Performance and Production*. Continuum Int. Publishing Group, 2003.

[130] Dirk Moelants, Olmo Cornelis, and Marc Leman. Exploring African tone scales. *Proc. 10th International Society for Music Information Retrieval Conference*, pages 489–494, 2009.

[131] Bob Monaco and James Riordan. *The Platinum Rainbow: How to Make it Big in the Music Business.* Omnibus, 1980.

[132] Daniel Müllensiefen. Fantastic: Feature ANalysis Technology Accessing STatistics (In a Corpus). Technical report, Goldsmiths, University of London, 2009.

[133] Daniel Mullensiefen, Joshua Fry, Rhiannon Jones, Sagar Jilka, Lauren Stewart, and Victoria J Williamson. Individual Dif-

ferences in Spontaneous Musical Imagery. *Music Perception*, 31(4):323–338, 2014.

[134] Daniel Müllensiefen and Andrea R. Halpern. The Role of Features and Context in Recognition of Novel Melodies. *Music Perception*, 31(5):418–435, 2014.

[135] Meinard Müller. Dynamic Time Warping. In *Information Retrieval for Music and Motion*, pages 69–84. Springer Berlin Heidelberg New York, 2007.

[136] Meinard Müller and Sebastian Ewert. Towards Timbre-Invariant Audio Features for Harmony-Based Music. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):649—-662, 2010.

[137] Meinard Müller, Peter Grosche, and Nanzhu Jiang. A Segment-based Fitness Measure for Capturing Repetitive Structures of Music Recordings. In *Proc. 12th International Society for Music Information Retrieval Conference*, pages 615–620, 2011.

[138] Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142, 2013.

[139] Sir Isaac Newton. *Opticks: a treatise of the reflexions, refractions, inflexions and colours of light*. The Royal Society, London, 1704.

[140] Joseph C. Nunes, Andrea Ordanini, and Francesca Valsesia. The power of repetition: repetitive lyrics in a song increase processing fluency and drive market success. *Journal of Consumer Psychology*, 25(2):187–199, 2014.

[141] Alexey Ozerov, Pierrick Philippe, Frédéric Bimbot, and Rémi 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.

[142] Elias Pampalk. A Matlab Toolbox To Compute Music Similarity From Audio. In *Proc. 5th International Symposium on Music Information Retrieval*, 2004.

[143] Ioannis Panagakis, Emmanouil Benetos, and Constantine Kotropoulos. Music genre classification: a multilinear approach. In *Proc. 9th International Society for Music Information Retrieval Conference*, 2008.

[144] Yannis Panagakis, Constantine Kotropoulos, and Gonzalo R. Arce. Music genre classification via sparse representations of auditory temporal modulations. *Proc. XVII European Signal ...*, 2009.

[145] Renato Panda, Bruno Rocha, and Rui Pedro Paiva. Dimensional music emotion recognition: Combining standard and melodic audio features. In *Proc. 10th International Symposium on Computer Music Multidisciplinary Research*, pages 1–11, 2013.

[146] Maria Panteli and Hendrik 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.

[147] Jouni Paulus. Music Structure Analysis by Finding Repeated Parts. In *Audio and Music Computing for Multimedia Workshop*, 2006.

[148] Jouni Paulus. Improving Markov Model-Based Music Piece Structure Labeling With Acoustic Information. In *Proc. 11th International Society for Music Information Retrieval Conference*, pages 303–308, 2010.

[149] Jouni Paulus and Anssi P. Klapuri. Labelling the structural parts of a music piece with markov models. In S Ystad, R Kronland-Martinet, and K Jensen, editors, *Proc. 5th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, volume 5493 LNCS, pages 166–176, Copenhagen, Denmark, 2009. Springer-Verlag Berlin Heidelberg.

[150] Jouni Paulus, Meinard Müller, and Anssi P. Klapuri. Audio-based music structure analysis. In *Proc. 11th International Society for Music Information Retrieval Conference*, pages 625–636, 2010.

[151] Marcus T. Pearce. *The construction and evaluation of statistical models of melodic structure in music perception and composition*. PhD thesis, City University London, 2005.

[152] Marcus T. Pearce and Martin Rohrmeier. Music Cognition and the Cognitive Sciences. *Topics in Cognitive Science*, 4(4):468–484, 2012.

[153] Marcus T. Pearce and Geraint A. Wiggins. Auditory expectation: the information dynamics of music perception and cognition. *Topics in cognitive science*, 4(4):625–52, oct 2012.

[154] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical report, Ircam, 2004.

[155] Geoffroy Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proc. 8th International Society for Music Information Retrieval Conference*, 2007.

[156] Geoffroy Peeters and Victor Bisot. Improving Music Structure Segmentation Using Lag-Priors. In *Proc. 15th International Society for Music Information Retrieval Conference*, pages 337–342, 2014.

[157] Carlos Silva Pereira, João Teixeira, Patrícia Figueiredo, João Xavier, São Luís Castro, and Elvira Brattico. Music and Emotions in the Brain: Familiarity Matters. *PLoS ONE*, 6(11), 2011.

[158] Isabelle Peretz, Dominique Vuvan, Marie-Élaine Lagrois, and Jorge L Armony. Neural overlap in processing music and speech. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370:20140090, 2015.

[159] R Plomp and W J Levelt. Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America*, 38(4):548–560, 1965.

[160] Graham E. Poliner and Dan Ellis. A Classification Approach to Music Transcription. *Proc. 6th International Society for Music Information Retrieval Conference*, 2005.

[161] Graham E. Poliner, Daniel P.W. Ellis, Andreas F. Ehmann, Emilia Gómez, Sebastian Streich, and Beesuan Ong. Melody Transcription From Music-Audio: Approaches and Evaluation. *IEEE Transactions on Audio Speech and Language Processing*, 15(4):1247, 2007.

[162] Jose Quesada. *Creating your own LSA space*, pages 71–85. Lawrence Erlbaum associates, Mahwah, New Jersey, 2007.

[163] Colin Raffel and Daniel P.W. Ellis. Large-Scale Content-Based Matching of Midi and Audio Files. *Proc. 16th International Society for Music Information Retrieval Conference*, pages 234–240, 2015.

[164] Roger Ratcliff. A theory of memory retrieval. *Psychological Review*, 85(2):59–108, 1978.

[165] Bruno Rocha, Niels Bogaards, and Aline Honingh. Segmentation and timbre-and rhythm-similarity in Electronic Dance Music. *Proc. Sound and Music Computing Conference*, pages 1–29, 2013.

[166] Pablo H. Rodriguez Zivic, Favio Shifres, and Guillermo a Cecchi. Perceptual basis of evolving Western musical styles. *Proceedings of the National Academy of Sciences of the United States of America*, 110(24):10034–8, jun 2013.

[167] Martin Rohrmeier and Thore Graepel. Comparing Feature-Based Models of Harmony. In *Proc. 9th International Symposium on Computer Music Modeling and Retrieval*, pages 357–370, 2012.

[168] Justin Salamon and Emilia Gomez. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, aug 2010.

[169] Justin Salamon, Bruno Rocha, and Emilia Gómez. Musical genre classification using melody features extracted from polyphonic music signals. In *Proc. International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[170] Patrick E. Savage and Steven Brown. Toward a new comparative musicology. *Analytical Approaches to World Music*, 2(2):148–197, 2013.

[171] Patrick E. Savage, Steven Brown, Emi Sakai, and Thomas 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*, 2015.

[172] Nicolas Scaringella, Giorgio Zoia, and Daniel Mlynek. Automatic genre classification of music content. *IEEE Signal Processing Magazine*, 23(2):133–141, 2006.

[173] E. Glenn Schellenberg and Christian 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.

[174] Joan Serrà. *Identification of Versions of the Same Musical Composition by Processing Audio Descriptions*. PhD thesis, Universitat Pompeu Fabra, 2011.

[175] Joan Serrà, Alvaro Corral, Marián Boguñá, Martín Haro, and Josep Ll Arcos. Measuring the evolution of contemporary western popular music. *Scientific reports*, 2:521, jan 2012.

[176] Joan Serrà and Emilia Gómez. Chroma binary similarity and local alignment applied to cover song identification. *Audio, Speech, and ...*, XX:1–14, 2008.

[177] Joan Serrà, Emilia Gómez, and Perfecto Herrera. Audio cover song identification and similarity: background, approaches, evaluation, and beyond. In *Advances in Music Information Retrieval*, pages 307–332. Springer, 2010.

[178] Joan Serrà, Xavier Serra, and Ralph G. Andrzejak. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9):093017, sep 2009.

[179] Xavier Serra, Michela Magas, Emmanouil Benetos, Magdalena Chudy, Simon Dixon, Arthur Flexer, Emilia Gómez, Fabien Gouyon, Perfecto Herrera, Sergi Jordà, Oscar Paytuvi, Geoffroy Peeters, Jan Schlüter, Hugues Vinet, and Gerard Widmer. *Roadmap for Music Information ReSearch*. The MIReS Consortium, 2013.

[180] Andrew J.R. Simpson, Gerard Roma, and Mark D. Plumbley. Deep Karaoke: Extracting Vocals from Musical Mixtures Using a Convolutional Deep Neural Network. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 429–436, 2015.

[181] Joren Six and Marc Leman. Panako: a scalable acoustic fingerprinting system handling time-scale and pitch modification. In *Proc. 15th International Society for Music Information Retrieval Conference*, 2014.

[182] Jordan B.L. Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J Stephen Downie. Design and creation of a large-scale database of structural annotations. In *Proc. 12th International Society for Music Information Retrieval Conference*, pages 555–560, 2011.

[183] Peverill Squire. Why the 1936 Literary Digest Poll Failed. *Public Opinion Quarterly*, 52(1):125–133, 1988.

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

[185] Bob L. Sturm. The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval. *Journal of New Music Research*, 43(2):147–172, 2014.

[186] Hiroko Terasawa, Malcolm Slaney, and Jonathan Berger. The thirteen colors of timbre. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 323–326, 2005.

[187] Don Traut. Simply Irresistible: recurring accent patterns as hooks in mainstream 1980s music. *Popular Music*, 24(1):57–77, 2005.

[188] George Tzanetakis and Perry R. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[189] George Tzanetakis and Perry Cook. Audio Information Retrieval (AIR) Tools. In *Proc. International Symposium on Music Information Retrieval*, 2000.

[190] Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary Detection in Music Structure Analysis Using Convolutional Neural Networks. *Proc. 14th International Society for Music Information Retrieval Conference*, 2014.

[191] Jan Van Balen, Dimitrios Bountouridis, Frans Wiering, and Remco Veltkamp. Cognition-inspired Descriptors for Scalable Cover Song Retrieval. In *Proc. 15th International Society for Music Information Retrieval Conference*, pages 379–384, Taipe, Taiwan, 2014.

[192] Jan Van Balen, Joan Serrà, and Martín Haro. Automatic Identification of Samples in Hip Hop Music. In *Int. Symp. on Computer Music Modeling and Retrieval (CMMR)*, London, United Kingdom, 2012.

[193] Aäron Van Den Oord, Sander Dieleman, and Benjamin Schrauwen. Transfer Learning by Supervised Pre-training for

Audio-based Music Classification. *Proc. 15th International Society for Music Information Retrieval Conference*, pages 29–34, 2014.

[194] Ralf van der Ham. Cover song detection : An evaluation for old popular music (unpublished thesis), 2014.

[195] Anja Volk and W. Bas de Haas. A Corpus-Based Study on Ragtime Syncopation. *Proc. 14th International Society for Music Information Retrieval Conference*, 2013.

[196] Anja Volk, W. Bas de Haas, and Peter Van Kranenburg. Towards Modelling Variation in Music as Foundation for Similarity. In *Proc. 12th International Conference on Music Perception and Cognition*, pages 1085–1094, 2012.

[197] Anja Volk and Peter van Kranenburg. Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae*, 0(0):1—-23, 2012.

[198] Thomas C. Walters, David A Ross, and Richard F Lyon. The Intervalgram : An Audio Feature for Large-scale Melody Recognition. In *Proc. 9th International Symposium on Computer Music Modeling and Retrieval*, pages 19–22, 2012.

[199] Avery Li-Chun Wang. An industrial strength audio search algorithm. In *Proc. 4th International Symposium on Music Information Retrieval*, 2003.

[200] Larry Wasserman. Undirected Graphical Models (Statistical Machine Learning, Lecture Notes), 2015.

[201] Frans Wiering. Balancing computational means and humanities ends in computational musicology (Utrecht University Humanities Lectures), 2012.

[202] Geraint A. Wiggins. Semantic Gap?? Schemantic Schmap!! Methodological Considerations in the Scientific Study of Music. *11th IEEE International Symposium on Multimedia*, pages 477–482, 2009.

[203] Geraint A. Wiggins, Daniel Müllensiefen, and Marcus T. Pearce. On the non-existence of music: Why music theory is a figment of the imagination. *Musicae Scientiae*, pages 231–255, 2010.

[204] Ian H. Witten and Darrell Conklin. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51—-73, 1995.

[205] Changsheng Xu, Namunu C. Maddage, and Mohan S. Kankanhalli. Automatic Structure Detection for Popular Music. *IEEE Multimedia*, 13, 2006.

[206] Yi-Hsuan Yang and Homer H. Chen. Machine Recognition of Music Emotion. *ACM Transactions on Intelligent Systems and Technology*, 3(3):1–30, 2012.

[207] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H. Chen. A Regression Approach to Music Emotion Recognition. *IEEE Transactions On Audio Speech And Language Processing*, 16(2):448–457, 2008.