Multifaceted Approaches to Music Information Retrieval

Cynthia C. S. Liem

Propositions

accompanying the dissertation

MULTIFACETED APPROACHES TO MUSIC INFORMATION RETRIEVAL

by

Cynthia Cheng Sien LIEM

1. Interpretation ambiguity of digital media items is an asset for search engine scenarios, rather than an undesired effect that should be factored out. [this thesis]
2. Narrative without the inclusion of musical terms is a sufficiently stable means for expressing music information queries. [this thesis]
3. Removing Dutch from the curriculum reduces the national industrial impact potential of future TU Delft engineers.
4. While the concept of the 'flipped classroom' gives clear role expectations at the side of the student, the expectations at the side of the teacher are ill-defined and risk passive teaching strategies.
5. Each academic workshop at a conference should contain at least one junior organizing member who still is in progress of obtaining a PhD degree.
6. When success is measured by employing key performance indicators, mediocre performance will be the default standard.
7. When aiming to strengthen valorization and impact opportunities for academic work, a silver tongue has more value than a golden demo.
8. A role model should be considered as a baseline rather than an upper limit.
9. Establishing women-only networking activities defeats the purpose of raising gender awareness.
10. 'Work-life balance' is a misnomer for 'balance between activities one likes to do or not'.

These propositions are regarded as opposable and defendable, and have been approved as such by the promotor prof. dr. A. Hanjalic.

Multifaceted Approaches to Music Information Retrieval

MULTIFACETED APPROACHES TO MUSIC INFORMATION RETRIEVAL

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op 26 november 2015 om 12.30 uur

door

Cynthia Cheng Sien LIEM

Master of Science in Media and Knowledge Engineering,
Master of Music in Classical Piano,
geboren te 's-Gravenhage, Nederland.

Dit proefschrift is goedgekeurd door de

promotor: Prof. dr. A. Hanjalic

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. A. Hanjalic,	Technische Universiteit Delft
Onafhankelijke leden:	
Prof. dr. G. J. P. M. Houben,	Technische Universiteit Delft
Prof. dr. H. de Ridder,	Technische Universiteit Delft
Prof. dr. R. C. Veltkamp,	Universiteit Utrecht
Prof. dr. E. Chew,	Queen Mary, University of London
Ao.univ.Prof. dr. A. Rauber,	Technische Universität Wien
Dr. T. C. Walters,	Google UK Ltd

Cynthia Liem is a recipient of the Google European Doctoral Fellowship in Multimedia, and the work presented in this thesis has been supported in part by this fellowship. In addition, the work in this thesis has partially been supported through the FP7 Project PHENICX (STREP project, ICT-2011.8.2 ICT for access to cultural resources, Grant Agreement No. 601166, 2013 – 2016).

Keywords: music information retrieval, multimedia information retrieval, music data processing, multimodality, multidisciplinarity, performance analysis, connotation, narrative, use context

Printed by: Ridderprint BV

Front & Back: Cynthia & Vera Liem, after 'Floating Apples' by Sandy (Nelson) Maynard. A full attribution and further notes can be found at the back of this thesis.

ISBN 978-94-6299-238-2

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

To my parents, my grandparents, and Vera.

CONTENTS

SUMMARY

Music is a multifaceted phenomenon: beyond addressing our auditory channel, the consumption of music triggers further senses. Also in creating and communicating music, multiple modalities are at play. Next to this, it allows for various ways of interpretation: the same musical piece can be performed in different valid ways, and audiences can in their turn have different reception and interpretation reactions towards music. Music is experienced in many different everyday contexts, which are not confined to direct performance and consumption of musical content alone: instead, music frequently is used to contextualize non-musical settings, ranging from audiovisual productions to special situations and events in social communities. Finally, music is a topic under study in many different research fields, ranging from the humanities and social sciences to natural sciences, and—with the advent of the digital age—in engineering as well.

In this thesis, we argue that the full potential of digital music data can only be unlocked when considering the multifaceted aspects as mentioned above. Adopting this view, we provide multiple novel studies and methods for problems in the Music Information Retrieval field: the dedicated research field established to deal with the creation of analysis, indexing and access mechanisms to digital music data.

A major part of the thesis is formed by novel methods to perform data-driven analyses of multiple recorded music performances. Proposing a top-down approaches investigating similarities and dissimilarities across a corpus of multiple performances of the same piece, we discuss how this information can be used to reveal varying amounts of artistic freedom over the timeline of a musical piece, initially focusing on the analysis of alignment patterns in piano performance. After this, we move to the underexplored field of comparative analysis of orchestral recordings, proposing how differences between orchestral renditions can further be visualized, explained and related to one another by adopting techniques borrowed from visual human face recognition techniques.

The other major part of the thesis considers the challenge of auto-suggesting suitable soundtracks for user-generated video. Building on thoughts in Musicology, Media Studies and Music Psychology, we propose a novel prototypical system which explicitly solicits the intended narrative for the video, and employs information from collaborative web resources to establish connotative connections to musical descriptors, followed by audiovisual reranking. To assess what features can relevantly be employed in search engine querying scenarios, we also further investigate what elements in free-form narrative descriptions invoked by production music are stable, revealing connections to linguistic event structure.

Further contributions of the thesis consist of extensive positioning of the newly proposed directions in relation to existing work, and known practical end-user stakeholder demands. As we will show, the paradigms and technical work proposed in this thesis managed to push significant steps forward in employing multimodality, allowing for various ways of interpretation and opening doors to viable and realistic multidisciplinary

approaches which are not solely driven by a technology push. Furthermore, ways to create concrete impact at the consumer experience side were paved, which can be more deeply acted upon in the near future.

SAMENVATTING

Muziek heeft vele facetten: buiten dat het ons auditieve kanaal aanspreekt, prikkelt muziekconsumptie ook onze verdere zintuigen. Ook wanneer muziek wordt gecreëerd en gecommuniceerd spelen verschillende modaliteiten een rol. Hiernaast staat muziek verschillende manieren van interpretatie toe: hetzelfde muzikale werk kan op verschillende geldige manieren worden uitgevoerd, en verschillende publieken hebben verschillende receptie- en interpretatiereacties op muziek. Muziek wordt verder ervaren in veel verschillende dagelijkse contexten, die niet beperkt zijn tot directe uitvoering en consumptie van geïsoleerde muziekinhoud: in plaats daarvan wordt muziek vaak gebruikt om niet-muzikale situaties te contextualiseren, van audiovisuele producties tot aan bijzondere aangelegenheden en gebeurtenissen in sociale gemeenschappen. Tot slot is muziek een onderwerp dat in veel verschillende onderzoeksgebieden is bestudeerd, van de geesteswetenschappen en sociale wetenschappen tot aan de natuurwetenschappen, en—met de opkomst van het digitale tijdperk—ook in de techniekwereld.

In deze dissertatie wordt beargumenteerd dat het volledige potentieel van digitale muziekgegevens alleen kan worden ontsloten wanneer de vele facetten als hierboven beschreven in acht worden genomen. Dit wereldbeeld in acht nemende, stellen we verschillende nieuwe studies en methoden voor gericht op problemen uit het 'Music Information Retrieval'-gebied: het onderzoeksgebied dat in leven geroepen is om zich met de ontwikkeling van mechanismen voor analyse, indexering en toegang tot digitale muziekgegevens bezig te houden.

Een belangrijke hoofdbijdrage van deze dissertatie wordt gevormd door nieuwe methoden om datagedreven analyses van meerdere opgenomen muziekuitvoeringen uit te voeren. De voorgestelde oplossingen zijn 'top-down' en houden zich bezig met het verkennen van overeenkomsten en verschillen binnen een corpus van meerdere uitvoeringen van eenzelfde werk. We bespreken hoe deze informatie gebruikt kan worden om wisselende hoeveelheden van artistieke vrijheid over de tijdlijn van een muzikaal werk te onthullen. Allereerst ligt de focus hierbij op de analyse van synchronisatiepatronen in piano-uitvoeringen. Hierna wordt een stap gemaakt naar het minder onderzochte gebied van vergelijkende analyse van orkestopnamen, waarbij we aangeven hoe verschillen tussen orkestuitvoeringen kunnen worden gevisualiseerd, uitgelegd en aan elkaar kunnen worden gerelateerd door technieken te gebruiken die oorspronkelijk waren voorgesteld voor visuele herkenning van menselijke gezichten.

De andere hoofdbijdrage van deze dissertatie gaat over de uitdaging om automatisch passende soundtracks te suggereren voor niet-professionele, gebruikersgegenereerde video. Voortbouwend op gedachten uit de Musicologie, Mediastudies en Muziekpsychologie stellen we een nieuw prototypisch systeem voor dat expliciet vraagt om het bedoelde verhaal (narratief) voor de video. Door informatie uit collaboratieve online informatiebronnen te gebruiken worden connotatieve verbindingen tussen dit verhaal en muzikale beschrijvingen gemaakt, gevolgd door herrangschikking op grond van audiovisu-

ele signaalanalyse. Verder verwerven we inzicht in wat voor narratieve eigenschappen op relevante wijze gebruikt kunnen worden bij het uitdrukken van (impliciete) muzikale informatiebehoeften in zoekmachines. Hiervoor bestuderen we welke elementen in vrije narratieve beschrijvingen, opgeroepen door het beluisteren van productiemuziek, stabiel zijn, waarbij verbindingen met taalkundige structuren voor uitdrukking van verschillende types gebeurtenissen worden onthuld.

Verdere bijdragen van de dissertatie bestaan uit een uitgebreide positionering van de nieuw voorgestelde onderzoeksrichtingen ten opzichte van bestaand werk, en bekende eisen en wensen van eindgebruikerspartijen. We demonstreren dat de paradigma's die in deze dissertatie met bijbehorend technisch werk zijn gepresenteerd erin zijn geslaagd om significante voortgang te creëren in gebruik van multimodaliteit, het toestaan van verschillende manieren van interpretatie, en het openen van deuren naar levensvatbare en realistische multidisciplinaire benaderingen die niet puur door een 'technology push' worden aangedreven. Hierbij is ook de weg vrijgemaakt naar mogelijkheden om concrete impact aan de kant van gebruikerservaring te krijgen, waar in de nabije toekomst actie op kan worden ondernomen.

INTRODUCTION

MUSIC BEYOND SOUND

To many people, music is considered to be an auditory affair. Indeed, when music is played, it manifests in the form of an acoustic signal targeted at human *listeners*. But beyond this, if music would purely be about listening, why do we still like to go to live concerts, where the audio and performance quality likely is less perfect than on our audio installation at home?

Immersing in the experience of a performance, physically moving along to music we like, jointly relating to music with our social peers, and seeing and feeling the entourage of a music event are some of the many factors that make a concert experience rich and worthwhile to attend [Melenhorst and Liem, 2015]. Clearly, in music consumption[1], **multiple modalities** are at play and the final **experience** of the music consumer is extremely important.

Also when dealing with the way music is being conceived and communicated before a performance, we can notice multiple modalities of importance. In many music performance traditions, when a music piece is conceived, it will not be communicated to its intended performers over the auditory channel, but rather in a notated form. In classical music, composers 'record' their music in the form of a score, consisting of symbolic notation. In jazz and popular music, while notation is usually less specified, outlines of music pieces are still notated and communicated in the form of lead sheets. There is no one-to-one mapping between information in a symbolic score and a performed rendition of the score. In fact, through performance, **multiple renditions** of the score are created by different musicians, which each will reflect **different readings and interpretations of the score**. Hence, musical notation does not encode the full scope of music information relevant to a piece—musicians will add information to this when performing. It is this added information which makes music pieces interesting for audiences to listen to. Another important notion is that there is no single right answer on what constitutes the 'best' musical interpretation. In fact, different interpretations which are equally valid and appreciated may perfectly co-exist for the same piece.

In the genre of traditional (world) music, symbolic notation is less prevalent, and oral transmission is the main way in which a song is preserved. Still, even here, music usually does not just involve abstract sound, but also a non-musical context coupled to it. Songs may go together with lyrics telling a narrative. Furthermore, with or without lyrics, the songs typically have a dedicated utility beyond the act of 'making and passively listening to music', such as accompanying ceremonies, enabling dance, communicating with others, and emphasizing collective and individual emotions.

[1] Here, 'consumption' is meant in a general sense (as 'a human being interacting with music in a receiving role') which is not necessarily economical.

Strong **connections between music and non-musical contexts** do not just hold within the traditional music domain: we frequently encounter these in everyday life. In many audiovisual productions, ranging from commercials to cinematic productions, combinations and associations between music, imagery and underlying narrative are actively exploited to influence the audience's perception and attitude towards the production and its underlying messages.

In conclusion, the **image of a multifaceted gemstone** seems an appropriate metaphor for music:

- there are many different sides constituting the 'appearance' of music (both in terms of relevant modalities, as well as relevant ways of interpretation);
- the different sides need to be 'polished' to make music more valuable to audiences (by composers and performers, through the creation of compositions and performance of multiple renditions);
- and just like a well-polished gemstone would most strongly be appreciated when displayed or worn in suitable entourages (and even be capable of improving general perception of these entourages), music will have similar positive experience effects on audiences when played and used in suitable contexts.

OPPORTUNITIES FOR DIGITAL MUSIC DATA

In the digital world, music increasingly is recorded, stored and consumed in digital forms as well. Digitization offers possibilities to explore, discover and access more and more diverse information than could be achieved in the analogue world. To ensure that improved and enhanced information access and consumption indeed can be achieved, digital music data should properly be represented, stored, organized, managed and accessed. For this, the creation of automated music data processing techniques which are capable of dealing with these requirements become a necessity. The research field aiming to address this challenge is broadly known as the *Music Information Retrieval* (Music-IR) field.

Given the notions on multifacetedness given above, it would make sense to consider digital music data as a **multimodal** phenomenon constituted by hybrid content, which allows **various ways of interpretation**, and plays an important role in the **consumer experience** when **embedded in various use contexts** [Liem et al., 2011b]. However, in existing work in Music-IR, traditionally the paradigm was adopted that music is monomodal. In case of recorded music, music frequently is equated to an audio signal, with the corresponding underlying assumption that all relevant information regarding music is (explicitly or implicitly) encoded in that audio signal. In recent years, the Music-IR field has started to move towards the inclusion of information from other modalities (e.g. combinations of audio and score, audio and text, audio and video and audio and sensor information), although monomodal approaches to music content still are prevalent in the field.

In general, the Music-IR field has historically preferred to focus on obtaining information from music data which tends to be as generalizable, objective and absolute as

possible (for example, analyzing at what timestamps physical audio events such as onsets occur, seeking exact matches of songs in databases, and devising new models of audio content feature representation). Beyond this, active ongoing work does exist dealing with aspects of music data which are not as strongly defined (for example, audio structure analysis, song similarity beyond the identity, ranging from opus and cover song detection to music genre classification, and music emotion analysis). Still, also in this type of work, the community tends to focus most strongly on commonalities across a corpus and methods for general music item labeling, thus not taking into account the aspect of various ways of interpretation.

Finally, the agenda of the Music-IR field so far has strongly been driven by *technology push* considerations, focusing on directions which from a technical viewpoint are interesting and feasible to address. At the same time, music is universally consumed, and as a consequence, a wide variety of audiences are interested in engaging with digital music, including non-technical audiences. So far, the Music-IR field did not play an active role in assessing how these audiences can truly benefit from advances in the field. This leads to considerable gaps between wishes, interests and expectations of non-technical music stakeholders and consumers, and those of Music-IR technologists [Liem et al., 2012], and many open questions on how Music-IR technology can effectively enhance the experience of music in various use contexts.

Finally, music is a cultural phenomenon, and as such, it has been a topic of study in the Humanities and Social Sciences long before the Music-IR field came into existence. In order to fully exploit the potential of digital music information access and consumption (and ease adoption beyond the technical domain), it would be useful to adopt a **multidisciplinary** (or even more strongly, an inter- or transdisciplinary) strategy, integrating insights and viewpoints from these various research fields throughout the creation process of sophisticated Music-IR technologies. Given differing interests and ways of communication in the various research fields, this is a non-trivial procedure which usually is omitted. Still, we believe it is of essential importance in making Music-IR truly impactful.

CONTRIBUTIONS OF THIS THESIS

In this thesis, we aim to act upon the opportunities which are offered by treating music in a multifaceted way. Most strongly, we will aim to **advance the notions of exploiting and enabling various ways of interpretation** in digital music retrieval. This is done by investigating **data-driven approaches and novel retrieval mechanisms** targeted at surfacing the more artistic, subjective and connotative aspects from digital music data and its surrounding contexts, which in traditional approaches would usually be factored out in favor of generalization. Our proposed solutions are explicitly multidisciplinary, not only building forth on technological insights and methods, but also taking into account related insights from the Humanities and Social Sciences (in particular Musicology, Media studies, Music Psychology, and Music Performance studies). At the same time, we connect Music-IR methods to questions and challenges posed in the broader *Multimedia Information Retrieval* community, which can be framed similarly to the Music-IR domain: a lot of data under study considers cultural and social artefacts, meant for consumption by broad audiences in varying use contexts.

As we will demonstrate throughout the thesis, the directions and connections explored in this thesis are highly novel. None of our contributions are incremental advances upon established problem domains; instead, with both the adopted paradigm and considered relevant information going beyond traditionally studied tasks, corpora and literature, we deal with young and challenging matter on which no established ground truth exists. Still, in the various thesis chapters, we will demonstrate the potential and new insights on digital music that our proposed methods can offer as thoroughly as possible, opening many doors to further development of more sophisticated digital analysis and access mechanisms to digital music, which are suitable for varying use contexts, while acknowledging the existence of varying interpretations and disciplinary viewpoints.

MULTI-, INTER- OR TRANSDISCIPLINARY?

In this introduction, so far, the term 'multidisciplinary' has been posed multiple times. While this term is used ambiguously in literature, in this thesis, we will follow the consensus as posed by Choi and Pak [2006]. Quoting from their work, the definitions and differences between 'multidisciplinarity', 'interdisciplinarity' and 'transdisciplinarity' are aggregated from existing literature as follows:

> *"Multidisciplinary, being the most basic level of involvement, refers to different (hence "multi") disciplines that are working on a problem in parallel or sequentially, and without challenging their disciplinary boundaries. Interdisciplinary brings about the reciprocal interaction between (hence "inter") disciplines, necessitating a blurring of disciplinary boundaries, in order to generate new common methodologies, perspectives, knowledge, or even new disciplines. Transdisciplinary involves scientists from different disciplines as well as nonscientists and other stakeholders and, through role release and role expansion, transcends (hence "trans") the disciplinary boundaries to look at the dynamics of whole systems in a holistic way."*

Following the definitions above, the work outlined in this thesis should be considered as 'multidisciplinary'. First of all, while borrowing from ideas from other disciplines, the thesis remains a Computer Science-oriented Music Information Retrieval contribution. Secondly, the work carried out in this thesis, being new, was not yet generated within the momentum of a larger consortium representing stakeholders from various disciplines with active involvement throughout the thesis progress.

Nonetheless, at the time of thesis completion, as will be outlined in the general Conclusions of this thesis, significant steps have actually been made to go beyond multidisciplinarity: active interdisciplinarity between the Music and Multimedia Information Retrieval disciplines has been achieved. Next to this, the momentum of a larger consortium with stakeholders from various disciplines is under active establishment – most concretely through the PHENICX project[2]).

[2] 'Performances as Highly Enriched aNd Interactive Concert eXperiences', FP7 STREP project, ICT-2011.8.2 ICT for access to cultural resources, Grant Agreement No. 601166, 2013 - 2016, `http://phenicx.upf.edu`. Accessed November 4, 2015.

THESIS OUTLINE

This thesis consists of three main parts:

I In Part I, we give an extensive literature-backed justification of the multifaceted approaches taken in this thesis, reflecting on challenges identified in the Music-IR and Multimedia Information Retrieval research communities, and considering practical experiences regarding technology adoption by various music stakeholders (musicologists, musicians, industrial parties).

II In Part II, acting upon the 'multi-interpretation' aspect of music, we consider the availability of multiple recorded renditions of the same musical piece. We discuss how this availability can be exploited by data analysis methods to gain more insight into interpretation aspects of the piece, which can be used for archive exploration use cases. In this, we consider the analysis of timing patterns in piano performance corpora, as well as the analysis of interpretation patterns in corpora of symphonic orchestra music through PCA-based spectrogram analysis.

III Finally, in Part III, we discuss the problem of suggesting soundtracks for user-generated videos, which was a 'Multimedia Grand Challenge' posed by Google at the 20th ACM International Conference on Multimedia[3]. In response to this challenge, we propose a prototypical system which makes use of information from collaborative web resources to automatically connect music to non-musical concepts. Next to this, we investigate more deeply how music can be connected to narrative user descriptions of intended multimedia contexts.

Each part starts with a short introductory overview, followed by a collection of related publications. The technical discussion will be finished by a general Conclusions chapter, reflecting on obtained achievements, open challenges, and future work.

PUBLICATIONS INCLUDED IN THIS THESIS

While a full publication list is given at the end of this thesis, the following publications will integrally be included as main chapters of this thesis:

8. **Cynthia C. S. Liem** and Alan Hanjalic. Comparative Analysis of Orchestral Performance Recordings: an Image-Based Approach. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 302–308, Málaga, Spain, October 2015. [Chapter 5]

7. **Cynthia C. S. Liem**. Mass Media Musical Meaning: Opportunities from the Collaborative Web. In *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 689–696, Plymouth, UK, June 2015. [Chapter 6]

[3] The ACM Multimedia Conference yearly features 'Multimedia Grand Challenges', in which industrial parties and other relevant stakeholders pose unsolved practical challenges in Multimedia, and research community members are invited to respond to these.

6. **Cynthia C. S. Liem**, Martha A. Larson, and Alan Hanjalic. When Music Makes a Scene — Characterizing Music in Multimedia Contexts via User Scene Descriptions. *International Journal of Multimedia Information Retrieval*, 2(1):15–30, 2013. [Chapter 7]

5. **Cynthia C. S. Liem**, Alessio Bazzica, and Alan Hanjalic. MuseSync: Standing on the shoulders of Hollywood. In *Proceedings of the 20th ACM International Conference on Multimedia — Multimedia Grand Challenge*, pages 1383–1384, Nara, Japan, October 2012. [Chapter 6]

4. **Cynthia C. S. Liem**, Andreas Rauber, Thomas Lidy, Richard Lewis, Christopher Raphael, Joshua D. Reiss, Tim Crawford, and Alan Hanjalic. Music Information Technology and Professional Stakeholder Audiences: Mind the Adoption Gap. In *Multimodal Music Processing*, Dagstuhl Follow-Ups vol. 3, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, pages 227–246, 2012. [Chapter 2]

3. **Cynthia C. S. Liem**, Meinard Müller, Douglas Eck, George Tzanetakis, and Alan Hanjalic. The Need for Music Information Retrieval with User-Centered and Multimodal Strategies. In *Proceedings of the 1st International ACM workshop on Music Information Retrieval with User-Centered and Multimodal Strategies (MIRUM) at ACM Multimedia*, pages 1–6, Scottsdale, USA, November 2011. [Chapter 1]

2. **Cynthia C. S. Liem** and Alan Hanjalic. Expressive Timing from Cross-Performance and Audio-based Alignment Patterns: An Extended Case Study. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 519–524, Miami, USA, October 2011. [Chapter 4]

1. **Cynthia C. S. Liem**, Alan Hanjalic, and Craig Stuart Sapp. Expressivity in Musical Timing in Relation to Musical Structure and Interpretation: A Cross-Performance, Audio-Based Approach. In *Proceedings of the 42nd International AES Conference on Semantic Audio*, pages 255–264, Ilmenau, Germany, July 2011. [Chapter 3]

I

SETTING THE SCENE: CASES FOR MULTIFACETED APPROACHES

OVERVIEW

As mentioned in the general Introduction to this thesis, throughout the thesis we consider music as a multifaceted phenomenon, and wish to push for multimodal and multidisciplinary approaches to Music Information Retrieval, which take various use contexts and audience types into account, and allow for various ways of interpretation of music content.

In this part of the thesis, we will set the scene by making a case for the proposed paradigms, discussing what the state-of-the-art in the field was at the beginning of the thesis project. The discussion is backed by literature and practical technology adoption experiences from various relevant stakeholder viewpoints.

The part consists of two chapters, which both are rooted in previously published work:

- In Chapter 1, we first pose the view of music as a multifaceted phenomenon in need of multimodal and user-centered approaches. Backed by literature, argumentation is given why such approaches are necessary to advance the field, and an overview is given of existing works along these directions. The work leading to this chapter was generated in the context of the establishment of the MIRUM workshop at the ACM Multimedia Conference; after two successful editions of this workshop, it was transformed into a dedicated 'Music, Speech and Audio Processing for Multimedia' submission area in the main technical track of the conference.

- In Chapter 2, we sketch practical experiences towards technology adoption with multiple stakeholder categories deemed relevant for the Music Information Retrieval domain. Music as a multimedia phenomenon only is relevant in the presence of human audiences, and Music Information Retrieval technology only becomes relevant if it will effectively manage to serve these human audiences. The work leading to this chapter was generated as a follow-up after discussions in a multidisciplinary group of experts and stakeholders at a Dagstuhl Symposium on Multimodal Music Processing. It reflects that achieving successful technology adoption is no trivial matter, while at the same time pinpointing the most common risks towards unsuccessful adoption.

The studies presented in this part have not only been important to the shaping of work presented in the remainder of this thesis, but also played a major role in the positioning and agenda establishment of the European PHENICX project, as will be discussed in the Conclusions chapter of this thesis.

1

THE NEED FOR MUSIC INFORMATION RETRIEVAL WITH USER-CENTERED AND MULTIMODAL STRATEGIES

Music is a widely enjoyed content type, existing in many multifaceted representations. With the digital information age, a lot of digitized music information has theoretically become available at the user's fingertips. However, the abundance of information is too large-scaled and too diverse to annotate, oversee and present in a consistent and human manner, motivating the development of automated Music Information Retrieval (Music-IR) techniques.

In this chapter, we encourage to consider music content beyond a monomodal audio signal and argue that Music-IR approaches with multimodal and user-centered strategies are necessary to serve real-life usage patterns and maintain and improve accessibility of digital music data. After discussing relevant existing work in these directions, we show that the field of Music-IR faces similar challenges as neighboring fields, and thus suggest opportunities for joint collaboration and mutual inspiration.

The contents of this chapter previously were published as Cynthia C. S. Liem, Meinard Müller, Douglas Eck, George Tzanetakis, and Alan Hanjalic. The Need for Music Information Retrieval with User-Centered and Multimodal Strategies. In *1st International ACM workshop on Music Information Retrieval with User-Centered and Multimodal Strategies (MIRUM) at ACM Multimedia*, pages 1–6, Scottsdale, USA, November 2011.

1

1.1. INTRODUCTION

Music is a universal phenomenon, which is studied, performed and enjoyed by a wide and diverse audience. Since the advent of the digital information age, music consumption has strongly shifted to digitized media and the World Wide Web, as is reflected in recent user surveys. A survey [Nielsen, 2011] involving 26,644 online consumers in 53 markets across the globe revealed that 'watching music videos on the computer' was the most broadly practised music consuming activity, done by 57% of these consumers in the 3 months before the survey was held. In terms of usage patterns, accessing digital tracks on the own computer is done the most frequently (49% several times a week, 28% daily), followed by the use of Internet video services, streaming services on the own computer, and social media sites. Another recent survey [Lidy and van der Linden, 2011] that was smaller in scale, but explicitly held among decision makers and stakeholders in the music industry, explicitly considered YouTube to be the number one music service, with iTunes only following after several personalized streaming services including last.fm, Pandora and Spotify.

YouTube is a video site and not a dedicated music service[1]. Nevertheless, people like to use it for consuming music, showing that the music experience is not just about listening anymore, but also about watching and sharing. The numbers associated with YouTube are impressive: overall, as of 2011 it has more than 3 billion views a day (nearly half of the world's population) and 48 hours of video uploaded every minute (amounting to seven years of video being uploaded per day)[2]. This means a large audience is within reach of an enormous diversity of multimedia data.

Music content is multifaceted, existing in many different representations. While originally being written down by a composer in the form of symbolic notation (e.g. in a score or a lead sheet), it usually only manifests when performed and presented to listeners in the form of music audio. Next to the symbolic and aural modality, multiple other modalities hold useful information that contribute to the way in which the music is conveyed and experienced: e.g. visual information from video clips and cover art, textual information from metadata, lyrics and background articles, and social community information on listening and rating behavior. This existence of complementary representations and information sources in multiple modalities makes music multimedia content, rather than a monomodal audio signal. Furthermore, the way music is experienced is strongly guided by affective and subjective context- and user-dependent factors.

In this chapter, we discuss why we consider the use of multimodal and user-centered strategies to be key to successful Music-IR solutions. We start with demonstrating that the nature of music is strongly connected to human factors, calling for a perspective on the content going beyond the audio signal. We show that this perspective is currently emerging, and discuss recent work on this, providing multiple relevant references. Finally, we compare current challenges in the Music-IR field to those encountered in neighboring fields, and argue that active collaboration between the Music-IR commu-

[1] Actually, since November 2014, YouTube has been offering music service facilities. Yet originally, the platform was not intended at all to serve this purpose.

[2] At the moment of publication of this thesis, no exact update numbers are available for 2015, but clear growth numbers are reported on `https://www.youtube.com/yt/press/statistics.html` (accessed November 4, 2015).

nity and these neighboring fields will be fruitful when addressing these challenges.

Due to space constraints, we omit an overall introduction to the Music-IR field here, but refer the interested reader to existing concise literature surveys in [Casey et al., 2008b, Downie, 2008, Downie et al., 2009, Kim et al., 2010, Müller, 2011, Orio, 2006, Paulus et al., 2010].

1.2. MUSIC GOES BEYOND AUDIO

1.2.1. A PHILOSOPHICAL PERSPECTIVE

As mentioned in the introduction, music exists in many different representations. In [Wiggins et al., 2010] it is suggested to categorize these into three domains that were originally mentioned by composer Milton Babbitt: (1) the *acoustic* or physical domain, (2) the *auditory* or perceived domain, and (3) the *graphemic* or notated domain. Each domain reflects aspects of music, but no single domain encompasses all of a musical object: in a certain sense, individual representations can be considered as projections of a musical object. The domains are connected through different types of transformations, as illustrated in Figure 1.1. In many Music-IR tasks, we are typically not interested in the precise (symbolic or digital) music encoding, nor in its sound wave dispersion behavior, but in the effect it has on human beings, which takes place in the largely black-boxed auditory domain. While music has communicative properties, it is not a natural language with referential semantics that indicate physically tangible objects in the world[3]. This poses evaluation challenges: a universal, uncompromising and objective ground truth is often nonexistent, and if it is there, there still are no obvious one-to-one mappings between signal aspects and perceived musical aspects. The best ground truth one can get is literally grounded: established from empirical observations and somehow agreed upon by multiple individuals.

1.2.2. MUSICAL SIMILARITY VS. AUDIO SIMILARITY

When audio content-based retrieval techniques in the Music-IR field started to develop, musical similarity was at first considered at two extreme ends of specificity [Casey et al., 2008b]. At the highest specificity level, the task of *audio identification* or *audio fingerprinting* (e.g. [Cano et al., 2002, Wang, 2003]) consists of identifying a particular audio recording within a given music collection using a small audio fragment as query input. Here, similarity is considered at the recording instance resolution. From a mathematical viewpoint, this type of similarity is close to the identity; in the context of the representation model in Figure 1.1, it deals with the graphemic and acoustic domain, but hardly with auditory aspects. While being robust towards noise, MP3 compression artifacts, and uniform temporal distortions, audio fingerprinting algorithms cannot deal with strong non-linear temporal distortions or with other musically motivated variations, such as the articulation or instrumentation.

At the other extreme end of the similarity spectrum, the task of audio-based *genre classification* (e.g. [Pampalk et al., 2005, Tzanetakis and Cook, 2002]) aims at assigning genre labels to songs. For this, spectral audio features with focus on timbre were heavily adopted, with Mel-Frequency Cepstral Coefficients (MFCCs) as a very popular repre-

[3] One can argue that lyrics can contain such information, but lyrics alone will not constitute music.

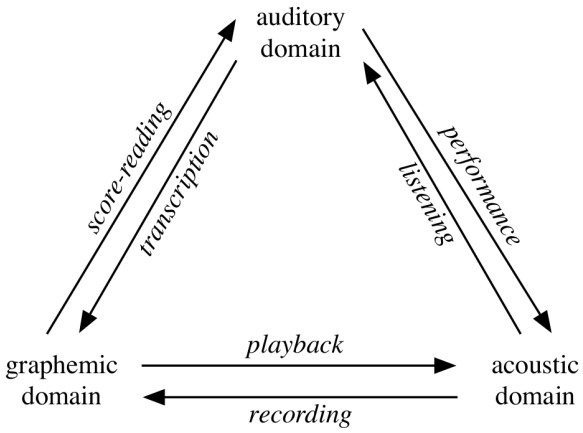


Figure 1.1: Three domains of music representations with transformations between them [Wiggins, 2009].

sentation. The adoption of MFCCs in music settings was mainly chosen because these features performed well in speech recognition settings, but were not motivated from a particularly musical or auditory perspective [Logan, 2000]. Extensive follow-up experiments (e.g. [Aucouturier and Pachet, 2004, Pampalk et al., 2005]) showed that timbral features were sensitive to production effects and that an apparent 'glass ceiling' was hit in terms of performance. In addition, in the context of music recommendation, collaborative filtering (CF) techniques matched or even outperformed audio content-based approaches [Barrington et al., 2009a, Celma, 2010, Slaney, 2011]. This seems to imply that either the adopted audio signal features were insufficiently comprehensive or appropriately modeled for the intended tasks, or audio signal information as a whole has been insufficiently comprehensive. Both of these hypotheses were shown to hold truth.

Regarding audio signal features, richer feature sets that e.g. incorporated rhythmic information next to spectral information led to improved classification results (e.g. [Pampalk et al., 2005, Pohle et al., 2009, Tsunoo et al., 2011]). In addition, attention shifted towards the mid-specificity level of music similarity. The task of *cover song identification* [Serrà et al., 2008] focuses on identifying different interpretations of the same underlying musical piece in a database, that can differ in instrumentation and harmony, represent different genres, or be remixes with a different musical structure. The closely related task of *audio matching* [Kurth and Müller, 2008] goes beyond identification at the *document level*, focusing on *fragment-level* matching of musically related excerpts in the documents of a given music collection. Tasks at this specificity level call for musically motivated audio features [Downie, 2008], and indeed have led to several successful feature representation proposals modeled on the concept of chroma [Gómez, 2006, Müller and Ewert, 2010, Müller et al., 2005]. A comprehensive recent overview of music-specific audio signal feature representations is given in [Müller et al., 2011].

The hypothesis that audio signal information does not account for all of the musical information is supported by findings in multiple Music-IR user studies. In the de-

sign process of a hypermedia library for recording inspirational ideas [Bainbridge et al., 2010], the foreseen musician users used spatial and visual means to record their ideas next to audio recording facilities. In [Aucouturier, 2009, Aucouturier et al., 2007] high-level semantic descriptions of songs were compared to acoustic similarity measures, and only weak mappings were found. An investigation of user query formulation [Lee, 2010] showed very associative notions of music, which are triggered by audio but ultimately are contextual, as e.g. can be seen in a real-life query description:

> *"I've heard the spooky tune, The Death March, several times tonight for Halloween. There are not words, just music. I've also heard the tune used in B-rated movies or cartoons to signify that someone or something has died. What is the origin of this tune? Who wrote it, when, and for what reason?"* [Lee, 2010]

In the context of recommender systems, artist graph analysis on MySpace showed little mutual information between the social and acoustic distance between artists [Fields, 2011]. In addition, users tend to rate and appreciate recommendations differently depending on whether they only get audio snippets provided or additional metadata [Barrington et al., 2009a].

This does not imply at all that audio content-based approaches do not use relevant information or should be abandoned. For example, as opposed to CF approaches, they do not suffer from a 'cold start' problem and thus can more robustly handle songs that were not (or infrequently) listened to before. Because of this, they yield more eclectic recommendations, while CF approaches will favor popular options [Celma, 2010]. Thus, audio content-based approaches are useful and should still be pursued and developed further—but for a more comprehensive perspective, holistic approaches should be taken that also take into account additional relevant information from other modalities.

1.3. MULTIMODAL AND USER-CENTERED MUSIC PROCESSING

Multimodal and user-centered[4] approaches to Music-IR tasks are already emerging in the community. While we will not be able to cover all of them in this chapter, we illustrate the current momentum in the field by giving several representative examples of current work in three broad topics:

- The combination of audio and textual data, which allows for automated and enriched tagging and classification of songs;
- Multimodal music synchronization, explicitly dealing with the temporal dimension of music;
- User-centered applications of multimodal strategies in interactive systems.

[4] We prefer to use the term 'user-centered' over 'user-centric', since this is lexically closer to 'user-centered design', a well-known Human-Computer Interaction approach implying that user aspects are practically considered from the start of the approach.

1.3.1. COMBINING AUDIO AND TEXTUAL MUSIC DATA

The by far most frequently adopted multimodal music approach combines audio data with textual (meta)data, which most commonly has been considered in the form of *web documents*, *lyrics* and *social tags*.

Web documents can reflect contextual, 'cultural' metadata information about music that is not trivially extractable from the audio signal, such as the country of origin of an artist and names of band members with their instrumental roles. In order to retrieve this information from the web, techniques have been proposed that either crawl the information from dedicated websites (e.g. last.fm, Wikipedia, audio blogs), or use the indexing results of a general-purpose search engine. Similarity as inferred from these techniques can be combined with audio-based similarity to allow for descriptive natural language querying of music search engines [Knees et al., 2007, Schedl, 2008, Whitman and Rifkin, 2002]. The strength of these cultural features, both as a monomodal feature set, as well as in combination with information from other modalities, has recently been underlined in [McKay et al., 2008].

Lyrics were mainly studied in connection to musical audio mood and genre classification. The performance of monomodal audio and lyrics feature sets was found to depend on the mood category [Hu and Downie, 2010a]. Multimodal approaches using both audio and lyrics features outperformed monomodal feature sets; as for the lyrics, the incorporation of stylistic features (e.g. rhyme, interjection words) increased performance in comparison to feature sets that only used a bag-of-words approach [Hu and Downie, 2010b, Mayer et al., 2008].

Social, community-contributed tags form a very important source of annotation information. These tags typically are obtained by means of a survey [Turnbull et al., 2008], through the harvesting of social tags [Eck et al., 2007], or by the deployment of annotation games [Law, 2011]. The acquired tags can be used to train audio-based autotaggers [Bertin-Mahieux et al., 2010]. In this context of automated annotation, a multimodal approach combining audio, social tag and web-mined data outperformed approaches using individual monomodal feature sets [Barrington et al., 2009b]. Community-contributed tags can be noisy, unstructured and very diverse; a recent approach to categorize them into multiple semantic facets is given in [Sordo et al., 2010].

Automated tagging has conventionally been applied at the document level, annotating a full song. However, the notion that tags can vary over time recently caught attention and work investigating this area is emerging [Mandel et al., 2010, Schmidt and Kim, 2010].

1.3.2. MULTIMODAL MUSIC SYNCHRONIZATION

The temporal dimension in music is reflected in many different music representations. When these representations are to be linked together in a multimodal setting (see Figure 1.2 for an illustration), it often is desired to link them over this temporal dimension, at the fragment level rather than the document level. For this, automated alignment or synchronization methods have to be applied. In this subsection, we will discuss two multimodal synchronization categories: *lyrics to audio*, and *audio to score*. A concise overview and discussion of synchronization techniques can be found in [Müller, 2007].

For lyrics to audio synchronization, several approaches have been proposed. In [Kan

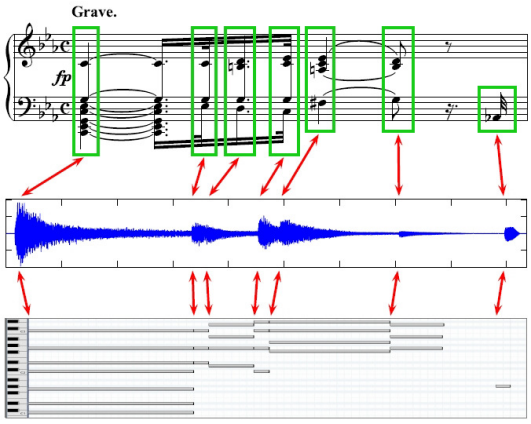


Figure 1.2: Linking structure (red arrows) over time of various representations of different modalities (sheet music, audio, MIDI) corresponding to the same piece of music [Müller et al., 2010a].

et al., 2008], structure analysis is performed on audio and lyrics separately, and a vocal detector is applied to the audio track. Subsequently, synchronization is obtained by aligning the found structural segments. An approach making stronger use of automated speech recognition techniques, including fricative detection and adaptation to the singer, is described in [Fujihara et al., 2011]. As an alternative to synchronizing lyrics and audio based on vocal feature detection, [Mauch et al., 2011] proposes a synchronization strategy employing chord annotations for lyrics in lead sheets, in combination with automated chord detection in the audio track.

Work on audio to score alignment started already three decades ago in the context of live score following for automated musical accompaniment, and developed ever since (e.g. [Dannenberg, 1984, Dannenberg and Raphael, 2006, Raphael, 2001a,b]). At present, both online and offline alignment techniques are being investigated. The online approaches, still being strongly geared towards interactive music performance systems, focus mainly on online anticipation of the musical timing of a performer [Cont, 2010, Raphael, 2010]. Offline approaches are geared towards alignment of audio tracks for indexing and navigation, and thus have more focus on temporal accuracy [Ewert et al., 2009] and scalability [Joder et al., 2011].

In general, audio to score alignment procedures have been developed assuming that the score is available in a digitized symbolic format, such as MIDI. In [Kurth et al., 2007], an approach was described that allows automated synchronization between scanned sheet music images and audio through optical music recognition.

1.3.3. MULTIMODAL AND INTERACTIVE SYSTEMS

The incorporation of multiple modalities and user-centered strategies already led to several successful interactive systems and prototypes. Much of the work described above is integrated in [Damm et al., 2012], where a digital library system is proposed for managing heterogeneous music collections. The player includes various document types, formats and modalities, and allows for high-quality audio playback with time-synchronous

Figure 1.3: Multimodal music player interfaces.

display of the digitized sheet music, playback of music videos, and seamless crossfading between multiple interpretations of the selected musical work. Figure 1.3 shows multimodal music player interfaces for this.

The challenge of visualizing music data collections and providing a user with means to navigate these collections in a personalized way is addressed in [Stober, 2011], where adaptive, multifaceted distance measures are being proposed. Several systems allowing for interactive manipulation of music content are presented in [Goto, 2011], where an outlook is also given on the future of music listening, which likely will shift more and more towards collaborative listening in social web communities. With wireless personal devices including sensors becoming increasingly common in daily live, collaborative music-making facilities that go beyond traditional instruments and making use of such devices are being developed [Ness and Tzanetakis, 2009, Zhou et al., 2011].

Regarding musical collaborative environments, very interesting work in a professional setting is described in [Zimmermann et al., 2008]. Here, distributed immersive performance is studied, focusing on ensemble playing facilities for musicians who are geographically at different locations. Experiments with the resulting multimodal system reveal how players use visual cues to establish synchronization, and show that latency tolerance is situation-dependent.

1.4. JOINT CHALLENGES, CROSS-DOMAIN OPPORTUNITIES

With the establishment of the International Symposium on Music-IR (ISMIR) in 2000, the Music-IR field became a focused community. Thus, the field still is relatively young in comparison to e.g. the Text-IR or Content-Based Image Retrieval (CBIR) fields.

At the 10th anniversary of ISMIR, an article was published looking back to the past years, and indicating directions and challenges for the near future [Downie et al., 2009]:

- Increased involvement of real end-users;
- Deeper understanding of the music data and employment of musically motivated approaches;
- Perspective broadening beyond 20th century Western popular music;
- The investigation of musical information outside of the audio domain;
- The creation of full-featured, multifaceted, robust and scalable Music-IR systems with helpful user interfaces.

It is striking how closely these resemble the directions suggested ten years earlier 'at the end of the early years' in CBIR [Smeulders et al., 2000]—with open questions that still have not been fully solved, as for example was seen in the closely related Multimedia-IR community [Hanjalic et al., 2008, Pavlidis, 2008]. Even in the Text-IR community, putting the user at the center of approaches is still considered a grand challenge rather than common practice [Belkin, 2008].

On one side, the Music-IR community can benefit from the many years of accumulated experience in neighboring communities, and transfer best practices from these fields to the own domain. Recent examples of this are the proposal of applying meta-evaluation as known in Text-IR to Music-IR evaluation [Urbano, 2011], and the transferring of scalable online image annotation to the music domain [Weston et al., 2011]. In addition, perceptually motivated sparse auditory features that were very successful in general sound retrieval settings recently attracted attention for applications in the music domain [Ness et al., 2011].

On the other side, the multifaceted, but non-referential nature of music data, and its strong connection to the human user, can lighten the path for other communities, and push research towards fundamental issues that currently still are not completely understood. The first successful cross-domain transfers in this way have already taken place. For example, the use of synthesized intermediate MIDI data for chord recognition model training formed the inspiration for an automated speech emotion recognition approach training on synthesized examples [Schuller and Burkhardt, 2010]. A successful music audio signal feature representation for audio matching settings [Kurth and Müller, 2008] formed the basis for an improved key-phrase detection approach in speech audio recordings [Zimmermann et al., 2008]. A cognitive model for music melody learning was shown to hold in a linguistics setting as well [Wiggins, 2011].

Finally, regarding user aspects, we consider the field of interactive music performance systems to have strong inspirational potential. Successful systems that manage to interact with humans, anticipate them and engage them in musical collaboration go a long way in terms of natural, non-verbal and grounded communication, and thus can deepen our understanding of successful human-computer interaction strategies.

1.5. CONCLUSION

In this chapter, we expressed the need for Music-IR with multimodal and user-centered strategies, outlined existing developments into these directions, and argued that music data in particular has the potential of addressing fundamental open issues that are largely encountered and unsolved in broad automated content-based settings. It is our hope that opportunities for cross-domain collaboration on these issues will be found and jointly explored soon.

2

ADOPTION CHALLENGES FOR MUSIC INFORMATION TECHNOLOGY

The academic Music Information Retrieval (Music-IR) discipline, which focuses on the processing and organization of digital music information, has multidisciplinary roots and interests. Thus, Music-IR technologies have the potential to have impact across disciplinary boundaries and to enhance the handling of music information in many different user communities. However, in practice, many Music-IR research agenda items appear to have a hard time leaving the lab in order to be widely adopted by their intended audiences. On one hand, this is because the Music-IR field still is relatively young, and technologies therefore need to mature. On the other hand, there may be deeper, more fundamental challenges with regard to the user audience. In this chapter, we discuss Music-IR technology adoption issues that were experienced with professional music stakeholders in audio mixing, performance, musicology and sales industry. Many of these stakeholders have mindsets and priorities that differ considerably from those of most Music-IR academics, influencing their reception of new Music-IR technology. We mention the major observed differences and their backgrounds, and argue that these are essential to be taken into account to allow for truly successful cross-disciplinary collaboration and technology adoption in Music-IR.

The contents of this chapter previously were published as Cynthia C. S. Liem, Andreas Rauber, Thomas Lidy, Richard Lewis, Christopher Raphael, Joshua D. Reiss, Tim Crawford, and Alan Hanjalic. Music Information Technology and Professional Stakeholder Audiences: Mind the Adoption Gap. In *Multimodal Music Processing*, Dagstuhl Follow-Ups vol. 3, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, pages 227–246, 2012.

2.1. INTRODUCTION

In the current digital era, technology has become increasingly influential in society and everyday life. This has led to considerable developments in techniques to process and organize digital information in many modalities, including sound. For the field of music, advancements have largely been geared towards two global goals: opening up new creative possibilities for artistic expression, and increasing (or maintaining) the accessibility and retrievability of music within potentially large data universes. Both of these goals additionally require attention for interaction opportunities, and may involve more modalities than mere sound. The academic field of research into these goals is typically characterized as *Music Information Retrieval* (Music-IR). This name was derived from *Information Retrieval*: a subdiscipline of computer science with applications in information (or library) sciences, employing established statistical techniques as a core component of its discourse, and most strongly focusing on textual data. Since a substantial amount of work in Music-IR actually does not actively deal with retrieval, the field has alternatively been called *Music Information Research*, retaining the same acronym.

The largest Music-IR success story so far may have been in audio fingerprinting (e.g. [Wang, 2003]), which is widely adopted in today's consumer devices[1]. Academic Music-IR research also unexpectedly found its way to a large audience through the Vocaloid[2] voice synthesis software, jointly developed by Yahama Corporation and the Pompeu Fabra university in Barcelona. Not long after the release of a voice package for a fictional character called 'Hatsune Miku', the character unexpectedly went viral in Japan, and now is also well-known to the Western audience because of her holographic concert performances, and her voicing of several Internet memes. Finally, through its API, the Echo Nest[3] powers multiple music-related applications that are reaching a broad audience.

However, for the rest, many of the academic Music-IR research agenda items apparently have a hard time leaving the lab to be successfully adopted in real systems used by real users. One can wonder if this is because the research field is too young, or if other factors are playing a role.

In business terminology, technological innovation can either be caused by *technology push*, in which new technology is internally conceived and developed to subsequently be 'pushed' into the market (while the market may not have identified an explicit need for it), or *market pull*, in which the research and development agenda is established because of an existing market demand. Initially, it may seem that the Music-IR research agenda is strongly driven by a pull: people need technology to keep overseeing the music information sources that they have access to, thus calling for fundamental and applied research advancements on this topic. But if this really would be the case, one would expect a much more eager adoption process, and a higher involvement of users and other stakeholders throughout the research process than encountered in daily practice.

When presenting envisioned new technology, and discussing their success potential

[1] It is not uncommon for an enthusiastic Music-IR researcher, trying to explain his research interests to a novice audience, to at one point get the question 'if he does something similar to Shazam', followed by a smartphone demonstration by the question-asker!

[2] `http://www.vocaloid.com`, accessed March 11, 2012.

[3] `http://the.echonest.com/`, accessed March 11, 2012.

with our academic peers, we typically assume that *some user already decided to adopt it*. In such a case, if user aspects are discussed (as e.g. is done in this Follow-Ups volume in [Schedl et al., 2012]), they will mainly concern strategies to optimize effective usage of the technology, giving the user a satisfying experience of it. The question *why* a user would want to adopt the technology in the first place is much less addressed and discussed at academic venues on Music-IR and related engineering disciplines; if it is, it is the realm of library science experts[4], not of engineers.

Of course, not every Music-IR research project has the urgence to immediately culminate into a monetized end-user system. Nonetheless, the Music-IR researcher will frequently have some prototypical beneficiary in mind. In several cases, this prototypical beneficiary professionally works with music (e.g. as a music sales person, producer, sound engineer, performing musician or musicologist), and the researcher will consider his Music-IR technology to be a novel and important enhancement to the daily practice of this music professional. However, it should be stressed that these envisioned professional music adopters do not typically come from the same backgrounds and mindsets as the academics who conceived the technology, and may actually not at all share the expectations of the academics regarding their work. Thus, involving this envisioned user, or even seeking fruitful academic collaboration with representatives of these user audiences, can prove to be much harder than expected.

Many authors of this chapter have shared backgrounds in both music information technology and professional music communities, or have worked closely with the latter. In this, it frequently was found that the successful embracement and adoption of new music technology by these communities cannot be considered an obvious, natural phenomenon that can immediately be taken for granted. In this chapter, we will share our experiences with this.

We will start by giving two concrete examples of systems that were created with a professional audience in mind, but received mixed responses. First of all, in Section 2.2, the reception of an intelligent audio mixing system is described. Section 2.3 will subsequently describe a case study on the *Music Plus One* musical accompaniment system, and discuss prevalent lines of thought in classical musicianship.

2.2. AUDIO MIXING

As a first example of how music technology was not received or adapted as expected by professionals, and a strong illustration of how sensitive the intended user can be, we will discuss the unexpected reception of an automated mixing system.

2.2.1. "IS THIS A JOKE?"

In the automatic mixing work of Enrique Perez Gonzalez and Joshua D. Reiss [Perez Gonzalez and Reiss, 2007, 2008a,b], intelligent systems were created that reproduce the mixing decisions of a skilled audio engineer with minimal or no human interaction. When the work was described in *New Scientist*, the response included outraged, vitriolic com-

[4] The library science field originally introduced the concept of *information needs*, a subject of study intended to justify or enhance the service provided by information institutions to their users. It includes topics such as information seeking behavior.

ments from professionals. Comments from well-known, established record producers included statements such as "Tremendously disappointed that you even thought this rubbish worth printing," "Is this a joke? Do these people know anything about handling sound," and "Ridiculous Waste Of Time And Research Budget[5]."

This reaction was surprising, since leaders in the field had previously expressed a desire and need for such research. For example, in his editorial for the *Sound on Sound* magazine of October 2008 [White, 2008], Paul White had stated that "there's no reason why a band recording using reasonably conventional instrumentation should not be EQed and balanced automatically by advanced DAW software." Similarly, James Moorer [Moorer, 2000] introduced the concept of an Intelligent Assistant, incorporating psychoacoustic models of loudness and audibility, to "take over the mundane aspects of music production, leaving the creative side to the professionals, where it belongs."

2.2.2. DIFFERING REACTIONS BETWEEN USER GROUPS

The hostility from practicing sound engineers and record producers may be due to several causes: a misunderstanding of the research, job insecurity due to fear of replacement by software, or simply a rejection of (and sense of insult from) the idea that some of their skills may be accomplished by intelligent systems. Of these causes, misunderstanding is quite plausible, despite the fact that the original article pointed out that the automatic mixing tools are "not intended to replace sound engineers. Instead, it should allow them to concentrate on more creative tasks." Other comments indeed revealed job insecurity: "I'm terrified because eventually this will work almost as good as someone who is "OK" and the cost savings will make it a necessity to most venue owners[6]." However, rejection of the idea that the technical skills of sound engineers and record producers might be automated is ironic, since music production already relies on a large number of tools that automate or simplify aspects of sound engineering, including acoustic feedback elimination, vocal riders and autotune.

Most interestingly, this negative reaction was not shared by musicians and hobbyists. One person's comments summed up the debate that occurred on many discussion forums: "I like this idea as a MUSICIAN, but not so much as a mixer. I've had so many shows I've played ruined by really bad sound mixers and seen so many shows that were ruined by a bad sound mix, that I welcome the idea[7]." Thus, it seems that people are comfortable with the idea of intelligent tools to address various aspects of music production and informatics, as long as those tools do not impact directly on their own.

Yet this attitude may be changed by providing the practitioners with demonstrations whereby they can experience first hand the effectiveness of new approaches. After a talk where audio examples of the automatic mixing research was presented, one professional audio engineer wrote "the power of automated mixing was effectively demonstrated – the result was perfectly reasonable for a monitor mix and, as the algorithms are per-

[5] This can for instance be seen on http://www.mpg.org.uk/members/114/blog_posts/190 and http://www.newscientist.com/article/dn18440-aural-perfection-without-the-sound-engineer.html, accessed March 11, 2012.

[6] http://thewombforums.com/showthread.php?t=14051, accessed March 11, 2012.

[7] http://www.gearslutz.com/board/so-much-gear-so-little-time/475252-software-company-begins-develop-program-replace-engineers-3.html, accessed March 11, 2012.

fected, the results will certainly improve further[8]."

The subsequent sections will deal with broader cross-disciplinary adoption and collaboration issues. For quite some time, Music-IR researchers have looked with interest to musicologists as a potential user audience. However, the amount of interest does not appear to be reciprocated, and Section 2.4 will elaborate on this, elucidating how current musicological interests are different from the common assumptions in Music-IR. Finally, a very different, but important category of professional users and collaboration partners is formed by stakeholders and representatives in the music industry. Section 2.5 will discuss current thinking and priorities for this audience, as voiced during the recent CHORUS+ *Think-Tank on the Future of Music Search, Access and Consumption.*

This chapter will be concluded with a discussion in Section 2.6, in which common adoption issues will be summarized and recommendations are given to overcome them.

2.3. PERFORMING MUSICIANSHIP

While the automated mixing system in the previous section was received well by musicians, a system that more closely approached music practice in a classical music setting has received varied responses by the intended user audience. In this section, experiences with the *Music Plus One* musical accompaniment system are described, with additional background information on classical music aesthetics that may (partially) explain the encountered reactions.

2.3.1. EXPERIMENTS WITH THE *Music Plus One* SYSTEM

For the last seven years, regular experiments have been performed with the *Music Plus One* musical accompaniment system, (a.k.a. the *Informatics Philharmonic*) [Raphael, 2001a,b, 2010], with students and faculty in the Jacobs School of Music at Indiana University. The program accompanies a musical soloist in a classical music setting with a flexible orchestral accompaniment that follows the live player and learns to do so better with practice. On the website of the system[9], the program can be seen in action. However, these videos only provide an 'external' view of the experience. The most important view of the experience is the soloist's: only the program's 'driver' will know how it responds, and how it manages to achieve the most high-level goal of allowing the soloist to become immersed in music making.

At this point, the author of the *Music Plus One* system has worked with over a hundred different soloists, including elementary school children, high school students, college players at all levels, as well as faculty. Most of the players are instrumentalists, with an emphasis on the strings, but also including wind and brass players. This group is not a cross-section of the classical music world, but rather represents an unusually dedicated and talented lot. For the most part, it is easy to convince young players to try out the computer as a musical partner. Most college level musicians also find the initial description of the experience appealing and are easily persuaded to bring their instruments to a rehearsal with the program. Before starting the experiments, it is first explained how

[8] `http://www.aes-uk.org/past-meeting-reports/intelligent-audio-editing-technologies/`, accessed March 11, 2012.

[9] `http://musicplusplus.net/info_phil_2011`, accessed March 11, 2012.

the computer differs from a human musical partner—the program's desire to follow the soloist might almost seem compulsive to a human musican, while it lacks a well-defined musical agenda of its own. Thus, the musicians are encouraged to be assertive and *lead* the performance; otherwise no one will.

Within a minute of playing it is usually possible to see how a musician will relate to the system. Some musicians never seem to take charge of the performance, mostly following the ensemble without asserting a strong musical agenda. However, most players immediately get the idea of leading the performance and are able to control the program simply by demonstrating their desired interpretation. While this inclination to lead is certainly correlated with the player's age, it has been interesting to observe how weak this association is. It is common both to have a talented 12-year old immediately getting the idea, while an occasional college player may never really catch on.

2.3.2. VERBAL AND NON-VERBAL RECEPTION FEEDBACK

What do musicians think of this program? While no formal or statistical approaches are adopted to measure their response, the sessions usually conclude with a brief discussion in which players share their thoughts. Most musicians that offer opinions are highly positive about the experience[10]. Many musicians say that it 'feels' like playing with a real orchestra and claim to find considerable enjoyment in the experience. In addition, many emphasize the value in preparing for 'real' performance.

However, the responses are not all positive. The most overtly negative reaction came from a composer on the faculty who had written an operatic scene for two voices and piano. Having heard about a public demonstration, he specificially requested a chance to try out the program. The situation was a particularly difficult one for the system, involving continually shifting tempo and mood, as is common in opera, along with the added difficulty of recognizing the voices. The composer criticized the program's lack of any internal musical agenda, placing (or misplacing) the desire to follow above all other musical considerations. In particular, he identified cases in which the timing of running notes in the piano was distorted for no apparent purpose, failing to create any natural sense of phrasing. This is a legitimate criticism, but it remains an open problem to even model the agenda of the accompanist, balancing an internal musical agenda with a desire to follow another musician.

Since actions speak louder than words, one might hope to gain a deeper understanding of players' attitudes toward the system by watching what they do, in addition to listening to what they say. In some ways, these actions have echoed the positive responses offered during the regular meetings. Several students have asked to use the program in their recitals, while the main faculty collaborator, professor of violin Mimi Zweig, has the program setup in her studio for use with her many students as an integral part of teaching. Judging from these examples, a certain degree of acceptance of this technology is observed.

On the other hand, it was routinely offered to students to give them the program, so that they can use it at home on their own computers. In spite of these many offers, only a few students have ever taken advantage of this offer. One particular graduate student

[10] Of course, it would be reasonable to expect that those who do not like the program may be more inclined to remain quiet.

comes to mind as typifying a common theme of response the program has received in the Jacobs School. She came to observe a prodigious young violinist practice with the system. The young violinist was considering the purchase of an expensive violin and had expressed interest in using *Music Plus One* to see how the instrument would project over an orchestra. The graduate student supervising this exchange was overwhelmed with excitement about the program's potential to make a lasting contribution to the classical musician. "This is going to change everything," she said.

Following this, numerous offers were made to the graduate student to rehearse with the orchestra or set the program up on her computer, though none ever materialized into any action on her part. Only indirectly it became clear that, while she saw the value the program had in an abstract sense, *she did not want to incorporate it into her musical world.*

2.3.3. CLASSICAL MUSIC VERSUS TECHNOLOGY: CONFLICTING OPPOSITES?

For a long time in Western history, music and mathematics were treated as close fields. In the ancient Greek era, philosophical writings described musical tuning systems together with their underlying mathematical ratios. In Mediaeval times, universities taught seven 'liberal arts': first the *trivium* consisting of grammar, logic and rhetoric, and afterwards the *quadrivium* consisting of geometry, arithmetic, astronomy—and music [Grout and Palisca, 2000].

However, this is not the image that many people would nowadays have of music. Instead, music is typically seen as a means of affective, personal expression, breaking through established formalisms and immersing the listener and player into a transcedent dream-like 'spirit' world, governed by emotion (and being far from the harsh, daily reality): a perspective that holds for classical music and popular music alike.

This perspective on music has its origins in the Romantic era. The notion of music being connected to emotional force had been acknowledged before: for example, the Baroque period strongly made use of musical formulae to express *affects*, a broad scala of human emotions. However, while the musical performer expressed the affects through his music, this mainly was a matter of rhetorical discourse, and he did not have to feel them himself, nor lead the listener into the affective states he was expressing.

The Romantic era brought new ideals, focusing on strong emotions, solitude, longing, and unreachable faraway realms. Ludwig van Beethoven lived and worked in the beginning of the Romantic era, and through his deafness, his seeming unwillingness to fit into society, and his (for that time) visionary and radical new music, many Romantic critics and writers considered him the prototypical Romantic Hero. This image of Beethoven as a suffering genius would dominate musical thinking for at least a century, and set an example for later generations. Performing musicians would mainly serve as servants to these composing geniuses, and each music listener attending a performance would experience the performance by getting lost in his own inner emotional world [Cook, 1998].

Such a Romantic aestethics perspective still is strongly represented in musical performance practice, at least for classical musicians. This may explain while for many generations, the classical music world has been rather resistant to new technology entering music practice (with the metronome, tuner, notation software as exceptions). Many musicians claim to greatly enjoy the experience of rehearsing with the computer, yet do not

want (yet?) to integrate a system like *Music Plus One* into their daily practice and teaching.

Informal discussions at another conservatoire gave similar outcomes regarding digital score material. While many musicians frequently consult the Petrucci Music Library[11] to check scores of potential repertoire, their attitude towards the possibility of digital music stands appears is ambivalent. While acknowledging the power of digital scores, several practitioners were opposed against using such a stand in a real concert performance, fearing that technology would let them down at a professionally critical moment.

Similar perspectives governed musicological thinking for a long time as well. However, present-day musicology has moved into more postmodern directions, and thus shows other adoption issues regarding Music-IR. These will be discussed in the following section.

2.4. MUSICOLOGY

Within the scientific Music-IR community, there is a strong but relatively informal agenda of advocacy to the musicological community of the tools and techniques being developed, often predicated on strengthening the case for developing the tools and on widening the areas of application in which they can prove their worth. It is not uncommon for keynotes at the ISMIR conference to raise the question of what Music-IR has to offer musicology, or how to attract musicologists to the field (e.g. [Cook, 2005, Downie, 2001, Downie et al., 2009]), and a musicologist at a Music-IR event can expect to be collared by any number of enthusiastic developers and asked "What would you like us to build?"

However, this advocacy seems often to fall on deaf ears; on the whole, musicologists do not seem to be adopting Music-IR techniques in their scholarship. The major journals of musicology rarely ever carry articles in which scholars have made use of computational techniques. For example, the *Journal of the American Musicological Society* (JAMS), which has seen 64 volumes up to the year 2011, includes just eleven articles which make oblique reference to computational subjects in its history[12]. In addition, very few undergraduate or graduate courses in music include teaching on computational methods: a non-exhaustive survey of course information from the USA, UK, Ireland, and Germany, published on the Web, reveals just six courses with explicit non-composition related computational components.

From a musicologists' point of view, it is easy to speculate on why computational Music-IR methods might not be eagerly adopted, beginning with the assumption that there are significant disciplinary, methodological, and scholarly discrepancies between (music) information retrieval and musicology. However, very little research has been carried out really attempting to give foundation to such speculation. This section focuses on the literature available on this topic, discussing discrepancies between the humanities and the sciences, mentioning practically encountered mismatches, and giving an outlook on how academic work in Music-IR and musicology can truly get closer to each

[11] `http://imslp.org/wiki`, accessed March 11, 2012.

[12] It should be noted that at least eighteen review articles in *JAMS* also mention computational subjects.

other.

2.4.1. MUSICOLOGY IN COMPUTATIONAL CONTEXTS: THOUGHT AND PRACTICE

Amongst the existing literature, a small number of studies have addressed questions regarding the information needs of musicologists, musicologists' use of recordings, and scholarly listening carried out in conjunction with a visualization. Brown [Brown, 2002] attempts to define the *research process* of musicologists using a variety of sociological research methods including semi-structured interviews and surveys. She found that, out of the stages of the research process she identified, the activity which musicologists value most highly is "keeping current" and also that they prefer journal browsing and face-to-face contact over digital communication to achieve this.

Although Cunningham's work [Cunningham et al., 2003] addressed more recreational information seeking, some of her conclusions are nevertheless relevant to scholars, particularly that advanced Music-IR techniques are not often developed beyond proof-of-concept into practical, usable tools.

Barthet and Dixon [Barthet and Dixon, 2011] conducted studies of musicologists examining performances using Sonic Visualiser[13]. They found that scholars were ambivalent towards the use of visualizations of sound. They appreciated that some timbral details were considerably more obvious in a visualization, but felt that timing and pitch details were much easier to hear than to see, and also that the visualization could distract listening in these cases.

While these studies may address some of the practical implications of doing musicology in a computational context, they do not address the discrepancies between the kinds of research carried out by the Music-IR and musicology research communities. For that, we may begin by turning to the inheritors of Charles Percy Snow, who postulated a fundamental divide in mindset between the arts (now more commonly referred to as the humanities) and the sciences in his now famous 1959 Rede Lecture, *The Two Cultures* [Snow, 1993], as well as the current of criticism in the digital humanities.

For example, Unsworth [Unsworth, 2005] highlights the perceived tension between *scholarship*, the often solitary, thought- and writing-directed process common in the humanities, and *research*, the often collaborative, problem-solving, question-answering, and hypothesis disproving process common in the sciences. For a musically specific example, Knopke and Jürgensen [Knopke and Jürgensen, 2011] claim as a benefit of computational music analysis that it is *consistent and repeatable*: features of research. However, the idea of *reproducibility* simply does not feature in contemporary music analysis. Musicologists do not see musical works as 'problems' requiring an analytical 'solution' which should be repeatable by other musicologists.

To give another example, Heinrich Schenker's theories on the workings of eighteenth and nineteenth-century Viennese music have a long tradition of being taken out of their context and codified as a universal method for uncovering fundamental structure in tonal music. However, Schenkerian analysis is not meant to yield one absolute truth, and should produce a subjective analysis unique to the analyst.

[13] Sonic Visualiser is a tool for interactive sound analysis providing a variety of visualizations, annotation, and plugin analytical procedures.

Unsworth also addresses the related concept of *systematization*, citing Northrop Frye who, in 1951, argued that "criticism"[14] ought to be systematic to distinguish it from other, less scholarly forms of cultural engagement. However, the idea of systematization is now treated with deep scepticism across the humanities, particularly in mainstream musicology. In general, since the second half of the 1980s, musicology has shifted into critical, postmodern directions [Cook, 1998], emphasizing subjectivity and cultural context, and refuting objective, universal, 'scientific' views on music.

2.4.2. A DISCIPLINARY DIVIDE

Another feature of this tension between humanities and computing is the status of technical contributions to humanities research. Many argue that interdisciplinary collaboration is the key to effective and credible technology adoption in humanities disciplines. However, Bradley [Bradley, 2009] argues that such collaborations are rarely considered as genuine equal scholarly partnerships. Rather, the technology is normally considered to be in the service of the scholarship and the partner from the humanities discipline is considered to be the "visionary", while "the technical person simply has the job of implementing the academic's vision." In this regard, the study of music represents a unique problem, since a technology-lead discipline focusing on music (Music-IR) exists *independently* of the humanities discipline (musicology).

Many scholars in the humanities generally focus on text as their source material, and are usually aware of the relative merits of computational approaches to working with text, such as the success of text search and the relative primitiveness of computational linguistics. By contrast, for music, the possibilities and limitations of dealing with the object of study (the music) tend to be less well understood or—to use an engineering term—harder. The complexity regarding the object of study (as e.g. outlined in [Wiggins, 2009]) has attracted scientists and technologists to cohere into a largely musicology-independent discipline, in which it is currently very common to see 'content-based' strategies being employed to approach music 'data'.

The indepency of Music-IR and musicology leads to a situation in which a lot of work that Music-IR researchers enthuse over is meaningless to musicologists. Frequently, the Music-IR technologists tend to focus on what seem, from a musicologist's perspective, to be more low-level 'problems' rather than higher level 'questions'.

A good example of this is the problem of *classification* which involves computational methods of determining properties such as 'genre' and 'mood' of examples of music from signal data. This is a challenging technical problem involving appropriate feature extraction and selection and testing of the statistical significance of results. But there is no equivalent question in musicology.

To make matters even more complicated, musicologists would often seek to problematize the kinds of genres which are routinely applied in Music-IR research. The meaning of 'genre' would already be questioned. In most musicological discourse, the term would mean something along the lines of the structural or compositional category of a musical work, such as 'symphony', 'string quartet', or 'song', while the kinds of labels

[14] Frye's use of the term "criticism" is taken from his background in literary studies. It is, in fact, the academic culture of literary criticism which really inspired much of the contemporary humanities, including musicology's re-invention as a critical discipline in the mid-1980s.

which Music-IR researchers apply as genre ('country', 'soul', 'funk', 'house', 'classical') would more likely be called something like 'style'. Musicologists would note the large number of styles missing from these lists ('renaissance vocal', 'lieder', 'acousmatic', 'serial', 'Inuit throat singing', etc.)—if accepting such lists at all, since the critical revolution in musicology has seen a rejection of the very idea of categorizing music into genres or styles at all.

Similarly, problems such as detecting the key or harmonic progressions in an audio signal require sophisticated computational approaches, but are the subject of undergraduate (or school-level) training in musicology, and would be taken for granted, or even not applied at all, at the level of professional scholarship. In fact, in British university music departments, technical competence in harmony and counterpoint and in aural analysis skills increasingly is diminishing in perceived importance[15]. In addition, automated analysis techniques of these types are not perfect yet and thus will make errors. This is very strange to a skilled expert, who may have to deal with ambiguities when making a manual analysis, but will never make such errors himself. If an automated technique will fail on very basic cases, its utility to the expert will thus be greatly reduced.

These examples begin to give an idea of the extent of the disconnect between these two approaches to music, and reasons why academics from one discipline who did not already have interest in the other discipline have not been eager to embrace work of this other discipline yet. The meeting point between mainstream thinking in these two disciplines is a great distance from each, and traversing that distance will require a considerable investment.

2.4.3. OUTLOOK FOR MUSICOLOGY

Since it seems that present-day musicology fundamentally has other interests than Music-IR researchers would initially assume, where does all this leave the advocates of Music-IR to musicology? One approach which is being taken is to introduce more musically sophisticated topics of research into the Music-IR agenda. Particularly, Wiering is encouraging investigation into the broad topic of musical meaning using Music-IR techniques [Wiering, 2009] and has, together with Volk, also been responsible for encouraging those working in Music-IR to find out more about contemporary musicology [Wiering and Volk, 2011], arguing that it is a "founding discipline" of Music-IR.

Looking the other way around, are there any aspects of the contemporary musicological research agenda which would suit computational techniques? One feature of the changes in musicology has been a shift of emphasis *away from musical works as autonomous objects*. A consequence of this is the study of *musical practice and its contexts*, including the study of performances and performers.

The study of musical performance provides a point of entry for audio-based computational techniques, i.e. a recording 'depicts' a performance (or possibly an edited amalgamation of several performances) and therefore provides a *handle on that performance as an object of study*. Amongst others, the Centre for the History and Analysis of

[15] This situation may be better for conservatoires, where these subjects are essential parts of the undergraduate (and sometimes even graduate) curricula in performing music disciplines. However, graduates of these disciplines are musicians, not musicologists.

Recorded Music[16] has been responsible for championing computational approaches to performance analysis in musicological contexts.

As an example, [Cook, 2007] uses a technique which analyses the differences and similarities in performance tempo and dynamics to infer genealogies of performer influence over a database of numerous performances of the same few Chopin *Mazurkas* over a period of around 70 years. An important difference between this study and a hypothetical identical study which would not make use of computers for the analysis is the *relative objectivity*. Here, the idea of consistency introduced above does becomes important, since in a study such as this, consistency of categorization of performance traits is vital for the credibility of the results. Perhaps the most fundamental difference, though, is that the hypothetical non-computational study is unlikely ever to have been conceived, let alone carried out: computational techniques afford scholarly investigations on a large-scale in a way which has never really been possible in the past, except by devoting a whole career to a project. The automated analysis of recorded performances also is being taken up in the Music-IR community already, e.g. in [Abesser et al., 2011, Grachten and Widmer, 2009, Liem and Hanjalic, 2011].

At a more global level, the interest of contemporary musicology in contexts around musical practice resonates very well with the current interest in Music-IR for multimodal and user-aware approaches—but this bridging opportunity has hardly been addressed or recognized yet. In addition, the situation that musicologists tend to problematize common assumptions, methods and vocabulary in Music-IR does not necessarily have to be a disadvantage. It can also open up new perspectives on situations that thus far were taken 'for granted' in Music-IR, but actually have not been fully solved yet.

2.5. MUSIC INDUSTRY: FINDINGS FROM THE CHORUS+ THINK-TANK

If the goal of an Music-IR researcher is to have his technology deployed and broadly adapted, stakeholders from music industry will often have to be involved. However, also for this category of collaboration partners, priorities and views on technology will differ.

In January 2011, MIDEM 2011, the world's largest music industry trade fair, was held in Cannes, France. At MIDEM, a *Think-Tank on the Future of Music Search, Access and Consumption* was organized by CHORUS+, a European Coordination Action on Audio-Visual Search[17]. Participation was by invitation only, limited to a small group of selected key players from the music and technology domains: highly qualified market and technology experts representing content holders, music services, mobile systems and researchers. In the months prior to the Think-Tank, an online survey about the future of the music business, music consumption, and the role of new technologies was held among opinion-leading decision makers and stakeholders across the music industry. Following the findings of this survey, the Think-Tank aimed at discussing current and future challenges of the music industry, and at assessing the role and impact of music search and recommendation technologies and services, including the latest developments from Music-IR research.

[16] CHARM, originally based at Royal Holloway, now at King's College London.
[17] http://avmediasearch.eu/, accessed January 28, 2012.

In this section, the findings of both the survey and Think-Tank roundtable discussions relevant to the topic of this chapter will be presented. The full report on the Think-Tank, as well as a full list of its participants, is available online [Lidy and van der Linden, 2011]. The participants who will feature in this section are Gerd Leonhard (CEO, The Futures Agency, MediaFuturist.com), Oscar Celma (Senior Research Engineer, Gracenote; formerly Chief Innovation Officer, BMAT), Rhett Ryder (COO, TheFilter.com), Stefan Baumschlager (Head Label Liaison, last.fm), Stephen Davies (Director Audio and Music, BBC), Holger Großmann (Head of Department Metadata, Fraunhofer IDMT), Gunnar Deutschmann (Sales Manager Media Network, arvato digital services), Laurence Le Ny (Music VP, Orange), Steffen Holly (CTO, AUPEO!), and Thomas Lidy (Founder and CEO, Spectralmind).

2.5.1. TRENDS AND WISHES ACCORDING TO STAKEHOLDERS

Gerd Leonhard was invited to give the keynote talk at the Think-Tank. In his presentation, he stressed the key changes in the music industry in the coming 3 to 5 years, all centered around one key word: *Disruption*. While participants of the survey agreed that the digital changeover had positive effects and that the digital music market has place for a wide range of diversified services, the digital changeover has been highly disruptive to the music business.

Consistent with other recent analyses, the survey named YouTube (which is actually not a music service!) as the number one music service. This popularity can e.g. be explained by the free access to the service, the presence of a broad and diverse collection[18]), the tendency of people not to change habits (i.e. platforms or services) frequently, and the added value of video.

The three main criteria for the choice of a music service were *availability of music*, *simplicity and 'ease of use'*, and *recommendation*. The emergence of streaming services seems prevalent, especially in the domain of music experts. Interestingly, this caused the more 'traditional' music service iTunes to be ranked in the survey *after personalized streaming services* such as last.fm, Pandora or Spotify and "other music streaming services / online radios", which were explicitly named by the participants.

According to the survey, the top five key enabling technologies for 2011–2020 will be *personalized recommendation*, *social recommendation*, *cloud services*, *audio-visual search* and *content-based recommendation*. In a follow-up free-form question, the following major trends for the future of music consumption were mentioned:

- instant availability and accessibility of music;
- automatic adaption of music to the (personal) environment, context;
- many ways of consuming music interactively;
- intuitive search, implicit search;
- personalization, unobtrusive recommendation;
- diversity, long-tail;

[18] While the collection is volatile and still subject to copyright claims!

- interoperability across services, global music profiles.

As a final question regarding technology directions, the survey participants were asked: "If a fairy granted you a wish for a technology (service, device ...) that would form the basis for a perfect product, what would you pick?". This led to the following wishes:

- "A (seamless and personalized) service that understands my current tastes, environment, mood and feelings, and can create for me a perfect stream of new music on the fly, wherever I am."
- "Play music for my current mood, play music to get me into a certain mood."
- "A music analysis system that analyzes the music not in objective terms but in terms of what a particular user will perceive."
- "An unlimited music streaming service with (cloud) locker capabilities, solid recommendations including long-tail coverage, social features to share music with friends and see what's trending with your friends; it should include additional artist info to explore biographies, pictures, recent news, tour info; it should have apps for all important smart phones."

The directions and wishes expressed above seem promising for Music-IR research, since they largely overlap with current academic research interests in Music-IR. However, it should be pointed out that many of the survey responders are expert opinion leaders with professional backgrounds in music technology. Thus, they form an 'early adopter' audience that may be stronger inclined towards new technological advances than 'the general public'.

2.5.2. PERSONALIZATION AND THE LONG TAIL

While contextual search, implicit search and multimodal forms of search were mentioned in the survey, personalized and social strategies were mentioned as the leading key enabling technologies for the future. Moreover, survey respondents stated that diversification and (recommendation of) non-mainstream content will be important to leverage music sales. At the same time, survey reponses showed that people mostly search for basic, specific and 'known' criteria, such as artist, composer, song title, album or genre. Apart from metadata-based search, other technology-enabled search possibilities such as search by taste, mood or similarity appear less prevalent. Discussions started about why this is the case: Because of no awareness that this is possible? Because the quality is not good enough yet? Or simply because there is no need?

The answer was two-fold: Oscar Celma suggested that the technologies are just not really in place yet. On the other hand, it was discussed under which circumstances such extended forms of search are really needed. Stephen Davies said consumers are quite simple in requirements: "Currently we put services that *we* think work. We need to better know what the users want." So is Music-IR research perhaps going in too complex directions regarding this?

Following the questions posted above, the role of the so-called *long tail* was discussed [Celma, 2010]. As Gerd Leonhard indicated, in the near future not music acquisition (or delivery) but *consumption* will be important. In a world where millions

of available music titles are available via streaming services, the main problem will be *choice*. Because of this, recommendation is important. However, in practice, only a tiny subset of the available content is seeing extreme usage, while the long tail beyond the popular artists is hardly consumed.

Music-IR technologies have a large potential to leverage the content in the long tail and make it (more) accessible. However, Gerd Leonhard stated that the problem is that most people will buy only what they know. Oscar Celma added, that 90 % of people are not very selective on music. Only a small percentage of enthusiasts really want content from the long tail; popular music is governing the choice of music.

Rhett Ryder reported that they inserted less known content from the long tail into the playlists at their service *TheFilter.com* and the acceptance was very high. This was confirmed by Stefan Baumschlager: Users desire new content—however, if it is too much, they will not like the service anymore. Thus, the right balance must be found between familiar and new content.

In Gerd Leonhard's opinion the long tail will not work unless the access is unlocked. Holger Großmann stated that most of the music portals do not offer mood-based or similarity-based search features yet. These technologies would give a different picture. Oscar Celma argued that for many services the clients are not the main goal, but making profits from the top artists. If only the top artists would matter, that would make the exploitation of the long tail through advanced Music-IR techniques no priority for industry.

Gunnar Deutschmann pointed out that exploiting the long tail will give an opportunity for small and independent artists. However, an open problem is how to get the music to the people. Music is frequently recommended personally by people, so it is unclear how to channelize the music to the audience.

As the survey participants already indicated, personalization will be important here. A successful music service should include recommendation based on user profiling, user feedback and deeper knowledge of the content, and usability and simplicity will be key factors for its success. These seem like very good arguments for the developments in Music-IR research. Yet, in order for them to be used by a large number of people, there still are issues to overcome, as will be discussed in the following subsection.

2.5.3. TECHNOLOGICAL OR BUSINESS MODEL ISSUES?

In some cases, research and development (R&D) in Music-IR technology has not matured enough yet to yield industry-ready tools. For example, Steffen Holly pointed out that the mixture and interaction of various technologies is not yet fully explored and that recommendation engines which combine various different criteria are key. Much more research on capturing and combining context information is needed (e.g. capturing the weather, combined with locations, and music playing in the car). Rhett Ryder added that all those factors and many more are important and need to be balanced correctly. Ideally, a device should be capable to capture and combine the sources of context independently of platform or service—although this will be a challenge on both the technological and business side.

In addition, there still are open research directions regarding *trust*. Stephen Davies mentioned that, since real personalization cannot be omitted, recommendation needs

to be based on trustable information (well-known DJs, etc.). This was also confirmed by Oscar Celma: recommendations from black-box machines give the user no trust, while friends' recommendations obtain much more trust. Recommendation engines need to give reasons for what they recommend.

However, for cases in which the necessary technologies are already there, the Think-Tank concluded that the main obstacles are *missing integration*, *unclear business models* and *legal issues*.

The basic technological 'bricks' for providing sophisticated music services do already exist: We have seen a tremendous growth of new music services around download, net radios, flat-rate based music streaming ('all access models'), new recommendation services, new technologies based on music analysis, music context and/or user profiling, personal radio based on collaborative filtering, etcetera. What is missing is integration: According to Laurence Le Ny the technological 'bricks' need to be integrated in a good way into a (global) music/entertainment universe and built on the right business model with easier access to rights and exhaustive offerings.

However, the business models are currently unclear[19]. There is concern about the wide availability of music ('why own something you can access for free on the Web?') and many startup companies struggle with rights issues around music licensing for the new consumption models. On the other hand, it should be easy to track music access and build business models and/or collection royalties on anonymized, proportional usage. In addition, Laurence Le Ny said the 'right' business model is not necessarily based on music alone but on a multi-screen personalised experience. She points towards a new simple and integrated music experience with different entry points and cross-media recommendation to cover consumers' needs and proposes bundling of services and offering subscription based models. However, she also points to difficulties in discussing these models with the majors in the music industry. Such business models take long to set up and require important negotiations with rights holders.

Finally, the question was raised if current business models leave margin for desirable Music-IR technologies at all. Steffen Holly said this is a big issue for recommendation technology providers. Content companies have to pay already a lot to collecting companies, licensing royalties, etcetera, making it very difficult to monetize a recommendation engine. Oscar Celma confirmed that it proves very difficult to sell a recommendation technology, even if it were the best in the world. Moreover, it is very difficult to communicate the added value around recommendations from the long tail. Holger Großmann agreed that there is no margin for these technologies in online stores. In the current business models new technologies cannot be paid, even if they are there and working already. A shift in monetization and royalty distribution is needed, but it is very difficult to achieve.

The Think-Tank participants agreed that the majors in music industry have a strong position but need to change in order to allow innovation. They also debated on the role of collection societies and the need for a shift from copyrights towards a public, open, standardized, non-discriminatory, collective, multi-lateral system of usage rights. The question is how to put all the stakeholders together in a common new business model.

[19] In fact, the highest disagreement in the survey was on the statement "Companies have clear strategies for revenue generation with digital music".

It is likely that changes in law and royalty distribution are needed. This is in line with the answers received from the survey on the major challenges to the (digital) music business, considering the number one challenge to be of legal/regulatory nature.

2.5.4. OUTLOOK FOR INDUSTRY

Current business models and legal issues seem to consider existing Music-IR technology to be sufficient for monetization purposes, and thus make it very hard for new and innovative Music-IR technology to get adopted. Does this mean that current Music-IR efforts are in vain?

Holger Großmann pointed to the need to distinguish between *recommendation* (main goal: selling) and *discovery services*. He believes that there is quite some space for R&D in the latter area. He mentions specific discovery scenarios: special content, searching sections within music, special business-to-business (B2B) use cases, etcetera. He also explains that as technology development is expensive, the rights holders must be prepared to share and to remunerate the technologist by some means or another. Oscar Celma said there is quite a market for search and discovery for professional users. There are also a number of specialized B2B markets, with specific use cases, such as production, sync, or the classical music market. This is confirmed by Thomas Lidy who experienced increasing awareness and interest in Music-IR technologies from production and broadcast areas in recent years.

As a conclusion, the discussions from the Think-Tank can be summarized as follows: many main technologies are there, but there is still room for research; R&D directions have been pointed out in the area of discovery. New services using more of the existing Music-IR technologies are expected to emerge, but business models still remain rather unclear.

A particular problem in this context are the adoption cycles of industry: Given that Music-IR technologies were not a priority of the industry for a long time, the take-up has been happening rather slow. Academic research meanwhile heads to new directions, not necessarily in line with the current needs of industry. Yet, the paradox is that the industry desires short innovation cycles and demands results to specific problems in short time.

A lot of research sees adoption only decades after its inception[20]. On the other hand, the market in the music domain is very fast-paced, and thus many times very simple solutions with no or little theoretical foundations are sufficient to appear on the market and have huge impact. These two different timelines—the fast-paced need for adopting solutions to stay ahead in the market versus the long time needed to obtain research results and elevate them to a mass-deployable solution—pose a significant challenge for the cooperation in research and development in this domain. This is complemented by an equally challenging legal situation that inhibits both research, by impeding the exchange of music data for collaboration and evaluation purposes, as well as deployment, with industry for a long time having been hesitant to adopt any solutions easing electronic access to music.

[20]Think of how long it took the vector space model and the concept of ranking in classical text IR to gain grounds on Boolean search, which still is a dominant search paradigm in many domains; or think of the time it took relational databases to catch foot in the mass market: long after the third normal form was invented.

As a good demonstrator of the potential impact and success of Music-IR research, there is a huge number of spin-offs created from PhD research in the field, many of which survive on the market, even gain huge value and are bought up by larger companies. However, we are faced with an environment for research and industry collaboration that offers a huge potential for R&D and real innovation, while at the same time posing rather severe constraints on its evolution.

2.6. DISCUSSION

In this chapter, multiple difficulties were pointed out regarding the adoption of Music-IR technologies by professional music stakeholders, and collaboration opportunities with these stakeholders towards the creation of such technologies. The main difficulties are summarized below.

FEAR OF REPLACING THE HUMAN

Users will not be inclined to adopt a technology if they feel threatened by it. In case of Music-IR technology, the technology may appear to threaten to replace the human in two ways. First of all, there is a perceived economical threat, in which the envisioned audience gets the impression that the presented technology will one day take over their daytime jobs. Secondly, there also can be a fundamental fear that technology takes over properties that were thought to be the unique domain of human souls: in this case, human musical creativity.

In both the audio mixing and music performing cases, it already explicitly was mentioned that it never has been the intention of the makers to 'replace' human beings with their technologies, but rather to provide ways to support and enhance sound producing and performing musicianship. This is a message that should remain to be emphasized.

From our case studies, it became clear that a 'not in my back yard' stance is realistic; while people recognize the use and benefit of new technology, they do not wish to have it entering their own professional and artistic worlds. It remains an open challenge on how to solve this problem; successful demonstrations by authorative early adopters appear still to be the best way, although a lot of patience will be necessary for this.

DIFFERING MEASURES OF SUCCESS

There may be mismatches regarding the notion of a successful system. While Music-IR inherited numeric success measures from the Information Retrieval field, measures such as Precision and Recall are often not convincing outside of these engineering communities.

In music performance applications, ease of use and a sense of naturalness in interaction will be a much more important factor. This did not just become clear for the *Music Plus One* system: in [Cont, 2011], describing the creative use of real-time score following systems, similar notions are made. For the task of real-time score following in an artistic context, *speed* will be more critical than *note-level accuracy*. In addition, in a musical creative context, the concept of *time* goes beyond discrete short-time low-level event detection: models are needed for higher-level temporal features such as tempo and event duration, and besides discrete events (e.g. pitch onsets), continuous events

(e.g. glissandi) in time exist too.

Care should be taken to identify the main goals of an intended user, since the user will be highly demanding regarding the capability of a new technology in reaching those goals. If expectations are not met, a system with new technology will be deemed immature and thus useless. For music performing and the creation of new music, as mentioned above, the rendering of an artistically convincing reaction to the user will be critical. In musicology, the concept of labeled 'truth' will be challenged. Even in industry, technology with high academic performance scores may not be useful if it does not fit the business model and does not allow for rapid monetization.

NEED FOR CONSIDERABLE TIME INVESTMENTS

Another important reason why Music-IR technology can face hesitance to be adopted has to do with the time required to achieve adoption. As was mentioned in the industry section, there is a strong mismatch between the deployment cycle timeline in industrial settings and the slower-paced academic research timeline, which has only become more delicate because of the late attention shift from industry towards digital music.

In addition, even cross-disciplinary collaboration needs considerable time investment to allow for serious and mutually equal cooperation between domains. Going back to the section on musicology, it takes time for musicologists to become familiar enough with tools and scholarly valid modes of discourse in information science and engineering – as it will take time for Music-IR scientists to become familiar with the scholarly valid modes of discourse and methodologies the other way around.

WRONG AUDIENCES?

In some cases, there might be unexpected other audiences for envisioned Music-IR tools. While the music industry stakeholders purely focusing on sales may not be interested in novel Music-IR technologies (or due to legal issues, not be able to consider them), stakeholders that rather focus on discovery aspects do allow for innovative R&D. While mid-level content-based analysis and classification systems hardly are of interest to the practice of musicologists, they can prove to be useful for performing musicians who prepare to study a piece. Finally, the postmodern interests of present-day musicology, with increased interest in subjectivity and contextual aspects, open up perspectives for multi- and cross-modal Music-IR research directions and linked data.

A striking feature of the Music-IR community is that many of its researchers do not just show affinity with research and the development of techniques to process their data, but that they are strongly engaged with the actual content of the data too. Both in- and outside their research, many Music-IR researchers are passionate about music and music-making[21]. For anyone working on new technology, but especially for people in this situation, it is important to be aware of realistic potential obstacles for the practical adoption of conceived technology.

This chapter was meant to increase awareness on this topic and to give a warning to the enthusiastic Music-IR researcher. As we demonstrated, several reception and adop-

[21] One could wonder if a similarly strong data engagement would e.g. hold for Information Retrieval academics and literature!

tion issues are of fundamental nature and may be very difficult to overcome.

On the other hand, this chapter was certainly not meant as a discouragement. There are many promising (and possibly unexpected) Music-IR opportunities to be found, that can lead to successful and enhanced handling of music information. However, in order to achieve this, careful consideration of the suitable presentation and mindset given the intended user audience, as well as investment in understanding the involved communities, will be essential.

II

DATA-DRIVEN ANALYSES OF MULTIPLE RECORDED MUSIC PERFORMANCES

OVERVIEW

In classical music, a musical piece typically is notated by a composer in the form of a score, and then performed by one or more musicians who usually are not this composer. Even more strongly, the same piece will typically be performed multiple times by multiple different performers, resulting in multiple readings and interpretations of the same notated material. As a consequence, when looking for recordings of a certain musical piece online, we often are offered a choice of many different versions.

From a search engine perspective, the availability of multiple performance versions for a musical piece can be considered as a near-duplicate problem. However, beyond resolving different performances to the same underlying musical piece, it also is interesting to consider the *similarities and differences* between various performances of the same piece, as these also carry important information. Audiences typically do not like to listen to deadpan computerized renditions of music scores. The 'added value' by a human performer considers shaping notated music information in such a way that it becomes understandable, expressive and interesting to the audience, which can be done by varying various musical parameters (among others, timing, articulation, balance and timbre). In doing this, the performer will not blindly randomize these parameters: certain places in the musical pieces allow for more freedom than other places.

In this part, we strive to surface performance conventions within corpora of multiple interpretations of the same piece, as implicitly encoded in the recordings constituting the corpus. By employing data-driven approaches, we aim to offer solutions which are scalable to larger and broader corpora without the need for annotation-intensive work connected to corpus changes, and ultimately move towards novel technology which can support the exploration and organization of multiple interpretations of a piece in a digitized archive.

This part consists of three chapters, which all are rooted in previously published work:

- In Chapter 3, we propose a novel method to gain insight into timing conventions in piano performances in the Mazurka corpus, based on standard deviation measures on automatically obtained alignment patterns between performances, and discuss how these automatic analyses can be related to structural analyses of the music pieces.
- In Chapter 4, we build forth on the method posed in Chapter 3, proposing entropy as an alternative to standard deviation in assessing alignment patterns, and presenting further quantitative and qualitative indications of how the obtained results are musically informative.
- In Chapter 5, we move away from the Mazurka corpus to the more challenging (and thus, underexplored) genre of orchestral performance recordings. With timing being a less distinguishing factor in performances in this genre, we adopt a

cross-disciplinary approach [Weninger et al., 2012] transferring the concept of eigenfaces in human face recognition to image-based spectrogram analyses of musical fragments, and show how these can be used to visualize differences between performers across a corpus.

3

A CROSS-PERFORMANCE, AUDIO-BASED APPROACH TO MUSICAL INTERPRETATION ANALYSIS

Notated musical scores provide a basic framework on which a performance is realized. Hidden behind this fixed score is a dynamic unwritten set of performance rules and conventions which performers draw upon when generating their own personal interpretation of the musical work. With the advent of recording, studying this implicit aural tradition is made possible through direct analysis of recordings. In this work, we exploit the availability of numerous recordings of the same piece and propose a lightweight, unsupervised and audio-based approach focusing on the analysis of expressive timing. In particular, we align multiple performances of the same piece by different performers and study timing deviations between aligned performances. This way, we gain insight into the degree of individualism in expressive musical timing displayed throughout a piece, which is part of the unwritten musical domain which performers add to the music. A qualitative study of the results of our technique applied to five Chopin mazurkas shows that timing individualism can be related to musical structure at both higher and lower structural levels, and highlights interpretation aspects that cannot objectively be found from a musical score only.

The contents of this chapter previously were published as Cynthia C. S. Liem, Alan Hanjalic, and Craig Stuart Sapp. Expressivity in Musical Timing in Relation to Musical Structure and Interpretation: A Cross-Performance, Audio-Based Approach. In *Proceedings of the 42nd International AES Conference on Semantic Audio*, pages 255–264, Ilmenau, Germany, July 2011.

3.1. INTRODUCTION

Many music pieces are originally conceived by composers and translated into scores. In order to be performed, the scores are studied and interpreted by musicians, who each give their own personal, expressive account of a score through their actual performance of the corresponding piece. Advances in recording technology as well as ongoing digitization efforts have caused that an increasing number of performances is available in digital form, both on the World Wide Web as well as in library archives. This way, we get access to an increasing number of different artistic readings of music pieces.

The way in which musicians manipulate sound has been compared to prosody: the way in which talkers manipulate speech in order to coordinate their discourse and clarify their expressive intentions [Palmer and Hutchins, 2006]. Musically expressive playing is formed through acoustical manipulations that can consider frequency, time, amplitude and timbre. These may go beyond the exact notation written down by a composer. Certain moments in a piece will allow for larger personal expressive freedom than other moments: in this, a performer is guided by theoretical rules, as well as stylistic traditions, personal taste and emotion, which are largely unwritten, dynamic and implicit aspects. Comparisons between multiple recordings of the same piece can provide information on common expressive patterns among different performers (such as tendencies to increase volume in ascending melodic lines, and slowing down at the end of musical phrases), but also highlight individual characteristics of a certain performer. This information can be used to reveal the interplay between performance aspects and the notated musical content: a relation which so far has not received a lot of attention in the Music Information Retrieval (Music-IR) field.

In this chapter, we consider expressive timing across different recordings of a piece. We assert that moments in the piece that show the largest timing deviations among performers, and thus the highest degree of individualism, must have given the performers a reason to do so, and therefore must be of a certain semantic relevance. Parallels can be found here with entropy in information theory, where items with the largest uncertainty regarding their actual realization contain the largest amount of information. Furthermore, moments showing a high degree of timing individualism are apparently not so obvious in interpretation that everyone would apply an identical timing strategy to them. This may clarify expressive content aspects of the piece itself that go beyond objective readings of a score, but still are essential for a successful musical performance.

Many existing studies on musical expression used expressivity parameters obtained from specialized recording devices, such as the Yamaha Disklavier. In the case of audio recordings, this data is usually not available. Because of this, we propose a fully audio-based method that does neither require external supervision, nor any intermediate high-level classification steps (e.g. explicit beat detection) throughout the procedure. Instead, we rely on the state-of-the-art work in audio-based music retrieval at the mid-specificity similarity level, and analyze alignments between music performance recordings in order to get insight into expressive timing deviations between performances throughout a piece. As such, our proposed method is intended to provide a lightweight supporting information source, that can enhance and aid several fields of Music-IR research. For example, in music-historical settings, the parts of a piece where the highest degrees of individualism are shown can be used for performer (or performer school) identifica-

tion and characterization. In automated music performance settings, the information can help in guiding the degree of expressivity in computer-rendered expressive performances throughout a piece, as a supporting channel to existing domain knowledge on expressive performance rules that typically is available in performance models. In automated retrieval and indexing settings, it is useful for identifying which parts of a recording will reflect the most typical characteristics of that recording in comparison to other recordings, allowing for quick and informed previewing in large databases.

This chapter is organized as follows. After a discussion of related work, our method for analyzing expressive timing deviations between performances will be described. This will be followed by a description of our experimental setup, and subsequently, by a discussion of the obtained results. After stating our main conclusions, the chapter will end with a discussion of future directions.

3.2. RELATED WORK

3.2.1. WORK USING AUDIO RECORDINGS OF MULTIPLE PERFORMANCES

Previous work exists in the Music-IR community, building on the availability of recordings of multiple performances of music pieces. First of all, in the field of audio-based music retrieval at the mid-specificity [Casey et al., 2008b] similarity level, work has focused on matching musically closely related fragments (*audio matching* [Müller, 2007, Müller et al., 2005]), or finding different versions of a song at the document level, which can range from different performances of the same notated score (*opus retrieval* [Casey et al., 2008b]) to potentially radically different new renditions of a previously recorded song (*cover song identification* [Serrà et al., 2008]). In general, matching and retrieval of classical music pieces has been shown to be achievable with near-perfect results [Casey et al., 2008a, Liem and Hanjalic, 2009].

Another category of previous work employing the availability of multiple performances has largely focused on the playing style analysis of individual performers (e.g. performer identification based on the way final *ritardandi* are played [Grachten and Widmer, 2009]). Furthermore, work has been devoted to visualization techniques to display performing style differences between different performers [Sapp, 2007, 2008], as part of a larger research project for which many classical music recordings (in particular of Chopin mazurkas) were collected, analyzed and annotated. However, in all these cases, no attention was paid to interpretational expressive aspects interacting with and possibly going beyond the notated music. With our work, we will make a novel step in this largely unstudied, but musically essential direction.

3.2.2. WORK ON MUSICAL EXPRESSIVITY AND RELATED MEANING

In Music-IR contexts, existing work dealing with musical expressivity largely had the modeling of general expressivity for computer-rendered expressive performances as goal (e.g. see [Goebl and Widmer, 2008, Goebl et al., 2007]). In the fields of music psychology and cognition, several studies have been devoted to the reasons for manifestation of musical expressivity. While the common assumption is that musical structure plays an important part in causing expressive playing, other causes have been mentioned as well, such as lower-level perceptual processes [Penel and Drake, 1998], as well

as alternate score editions, performer-specific treatments of ornamentation and pedaling, and music-theoretic notions of expectancy and tension-relaxation [Palmer, 1996]. In [Juslin, 2003], a model is proposed decomposing causes for musical expressivity into generative rules, emotional expression, random variability, motion principles and stylistic unexpectedness. Regarding motion principles, performers typically were observed to slow down at the end of phrases (phrase-final lengthening) [Palmer, 1996]. In [Friberg and Sundberg, 1999], models for the slowing down in the final *ritardando*, marking the end of a musical piece, were shown to be similar to the kinematic slowdown observed for decelerating runners.

A lot of work has focused on expressivity in piano playing (e.g. [Desain and Honing, 1994, Palmer, 1996, Penel and Drake, 1998, Repp, 1994]), probably because advanced facilities exist for recording expressivity in piano performance (e.g. Yamaha Disklavier, Bösendorfer SE), and the amount of expressive means in a piano, given its discrete tone set and percussive mechanics, is restricted to timing, dynamics and articulation. With regard to expressive timing, inter-onset-intervals (IOIs) are commonly used as features, which express the time between subsequent onsets. However, in the case of audio recordings without explicit expressivity data recorded at the time of the performance, such expressive features will have to be extracted from the audio signal. So far, it has shown to be very hard to do this automatically in an accurate way. As a compromise, semi-supervised methods are often applied, in which the results of initial automated methods are afterwards manually corrected (e.g. [Grachten and Widmer, 2009], but such an approach is infeasible for large data sets or long musical excerpts), or in which symbolic information, such as MIDI files of the performed piece, is used to guide the automated algorithms (e.g. [Müller et al., 2009]). In our approach, we will employ a fully unsupervised method, and also circumvent any high-level intermediate classification steps (e.g. beat or onset detection) throughout our procedure.

3.3. AUDIO-BASED ALIGNMENT OF MULTIPLE PERFORMANCES

As mentioned in the previous section, when only having an audio recording of a musical performance, aspects of expressivity will have to be extracted from the signal. In contrast to previous approaches, we propose to infer timing expressivity in an objective and data-driven way: by relying on state-of-the-art analysis and alignment procedures from audio-based music retrieval at the mid-specificity similarity level, and subsequently analyzing the alignment differences between multiple performances. Domain knowledge will be used in the interpretation and discussion of our results, but not in the analysis procedures themselves. Our method can be summarized as follows:

1. Let A be a set, containing n audio recordings of the same piece. For each available recording $a \in A$, short-time chroma-derived features r are computed that express the harmonic content of the recording at each short-time instance. The result is a set R with n elements $r \in R$. Each element r is a vector containing short-time chroma-derived feature profiles for a uniquely corresponding audio recording $a \in A$.

2. Choose a $r_{ref} \in R$ corresponding to the short-time feature vector of an arbitrary reference recording, and align the feature vectors of all other recordings

$r \in R \backslash \{r_{ref}\}$ with r_{ref}, obtaining $n-1$ alignment paths $w \in W$.

3. The alignment paths are further processed in order to emphasize transitional, irregular alignment behavior over regular behavior. Subsequently, at each short-time instance $t = 1 \ldots m$ we calculate the standard deviations over the full path ensemble represented by W, thus obtaining indications of between-performance variance corresponding with each short-time time instance of r_{ref}.

The first two steps, applying state-of-the-art techniques from audio-based music retrieval, are discussed in this section. Our main contribution in this chapter, which consists of analyzing the results of the alignment procedure, will subsequently be presented in Section 3.4.

3.3.1. AUDIO FEATURES

As short-time harmonic descriptor features, we adopt the state-of-the-art Chroma Discrete Cosine Transform-reduced Log Pitch (CRP) features [Müller and Ewert, 2010], that have been shown to outperform traditional chroma representations in terms of timbre-robustness and audio matching performance. In their design, these features combine techniques from classical chroma and Mel-Frequency Cepstral Coefficient (MFCC) features.

First of all, the audio signal is decomposed into 88 frequency bands corresponding to the MIDI pitches of a standard piano. Subsequently, the local energy for each (squared) frequency subband is expressed through the short-time mean-square power. After these steps, which are typical to traditional chroma computation techniques, an approach will be followed with strong similarities to the computation of MFCCs. The pitch energies first are logarithmically compressed to account for the logarithmic sensation of sound intensity in human perception. Afterwards, a discrete cosine transform (DCT) is performed, yielding Pitch-Frequency Cepstral Coefficients (PFCCs) with an interpretation analogous to that of MFCC features. In typical audio matching settings, matches between similar musical content is sought regardless of timbre. With the lower coefficients of MFCCs being considered to relate to timbre, a similar consideration holds for the PFCCs. Therefore, in order to emphasize timbre-invariance, the lower coefficients of the PFCCs are set to zero. Subsequently, after performing an inverse DCT, an enhanced pitch energy representation is obtained, which finally is projected onto the 12 chroma bins in order to obtain the CRP features.

We adopt the implementation of the CRP features as made available by the authors[1]. The default parameters are used, except for the short-time window length, which is shortened to an 2048-point Hamming window. With 50% overlap and the default sampling frequency of 22050 Hz, this means that short-time CRP profiles in any $r \in R$ will be computed every 0.0464 s. For more details on the CRP features, the reader is referred to [Müller and Ewert, 2010] and the publicly available implementation.

3.3.2. ALIGNMENT STRATEGIES

Different performances of the same piece may substantially differ in global tempo. As the feature computation approach described above uses short-time feature sampling

[1] `http://www.mpi-inf.mpg.de/~mmueller/chromatoolbox/`, accessed May 31, 2011.

with a fixed-window length, the resulting CRP profile vectors $r \in R$ will therefore be of different lengths. As internal timing differences will occur among performances, these vectors cannot be linearly scaled over the time dimension in order to make them comparable with each other. Instead, alignment methods will be necessary to find a time mapping between corresponding events in different recordings. This can be done through Dynamic Time Warping (DTW) methods.

In the field of cover song retrieval, a very powerful alignment technique was proposed in [Serrà et al., 2008], which we implemented from the literature. Here, a binary cost measure was used to express the match between two short-time profile vectors considered for alignment. In order to prevent pathological warpings, local constraints are imposed with respect to the alignment path.

We choose a CRP profile vector $r_{ref} \in R$, corresponding to a reference recording that may be arbitrary chosen. By aligning the vectors $r \in R \backslash \{r_{ref}\}$, corresponding to all other recordings in the set, with r_{ref}, full alignment between performances is achieved through this reference recording vector. For each alignment between r_{ref} and an $r \in R$, an alignment matrix X is constructed. The alignment value $X_{i,j}$ between two CRP profiles at time instances i and j in r_{ref} and r, respectively ($r_{ref}[i]$ and $r[j]$), is computed adopting the following local constraints:

$$X_{i,j} = \max \begin{cases} X_{i-1,j-1} + S_{i-1,j-1} - \phi(S_{i-2,j-2}, S_{i-1,j-1}) \\ X_{i-2,j-1} + S_{i-1,j-1} - \phi(S_{i-3,j-2}, S_{i-1,j-1}) \\ X_{i-1,j-2} + S_{i-1,j-1} - \phi(S_{i-2,j-3}, S_{i-1,j-1}) \\ 0 \end{cases}$$

where S is a matrix indicating the binary similarity values between all pairs of CRP profiles from r_{ref} and r, and ϕ is a penalty function:

$$\phi(a,b) = \begin{cases} 0 & \text{if } b > 0 \text{ (no gap)} \\ 0.5 & \text{if } b \leq 0 \text{ and } a > 0 \text{ (gap opening)} \\ 0.7 & \text{if } b \leq 0 \text{ and } a \leq 0 \text{ (gap extension).} \end{cases}$$

The gap opening and extension penalty constants were suggested in [Serrà et al., 2008]. Initialization procedures, binary similarity measures and other parameters were also taken from this article, to which the interested reader is referred for further technical details.

For cover song retrieval purposes, obtaining an alignment score from the DTW procedure is sufficient, but in our case, we need an explicit alignment path, that can be obtained by tracing back from the point corresponding to the highest total alignment score. If $|r_{ref}| = m$, for each alignment of r_{ref} with a performance short-time CRP profile vector r, we want to obtain an alignment path w of length m, with $w[1 \ldots m]$ indicating short-time instance indices of the CRP profiles in r that correspond to the CRP short-time instances found at $r_{ref}[1 \ldots m]$. Not all time instances $1 \ldots m$ may have been explicitly covered in the original alignment path, but assuming linear development for unknown instances, we can simply achieve this through linear interpolation.

3.4. PERFORMANCE ALIGNMENT ANALYSIS

After calculating all alignment paths following the procedures above, we will have obtained a set W with $n-1$ alignment paths $w \in W$, each of length m. We post-process these paths to emphasize irregular alignment behavior: if an alignment subpath $w[k \dots l]$ shows constant alignment steps, this means that the corresponding CRP profile feature subsequence in r is a linearly scaled version of $r_{ref}[k \dots l]$, and therefore does not reflect any timing individualism. In order to emphasize this timing individualism, we will therefore highlight alignment step slope changes, which is done by computing discrete second derivatives over the alignment paths.

First of all, for each alignment path w, we compute the discrete first derivative δ through the central difference:

$$\delta[i] = \begin{cases} \frac{1}{2}(w[i+1] - w[i-1]) & 1 < i < m \\ w[1] - w[0] & i = 1 \\ w[m] - w[m-1] & i = m. \end{cases}$$

With the path having been traced back, the imposed local alignment constraints and the initialization procedure of the alignment matrix, a large 'startup' derivative is typically found at the beginning of the path due to an initial alignment jump. As we are only interested in the alignment step development within the true alignment path (and the beginning of the recording for the given time sampling rate will contain silence), we set the derivative values up to this alignment startup point to zero. Subsequently, we repeat the central difference procedure on the enhanced δ to get a second derivative approximation δ^2. For each alignment path w, we compute this second derivative approximation vector δ^2, constituting a set Δ^2. Finally, for each time instance $t = 1 \dots m$, we compute the standard deviation over all alignment path second derivatives in Δ^2 at that time instance, obtaining a standard deviation curve of length m that expresses timing individualism over the full performance ensemble at moments corresponding with each short-time time instance of r_{ref}.

As argued before, the second derivative approach is locally time-scale invariant. On purpose, we do not impose any global time-scale invariances, as there are conflicting reports whether relational invariance of expressive microstructure holds under global tempo changes [Desain and Honing, 1994, Repp, 1994].

3.5. EXPERIMENTAL SETUP

3.5.1. DATA

A suitable dataset for our experiments, which contains a lot of performance recordings of individual musical pieces information, has been created as part of the Mazurka Project, undertaken at the AHRC Centre for the History and Analysis of Recorded Music (CHARM). For this project, many recordings (almost 3000) of Chopin mazurkas were collected and analyzed. For five of the mazurkas, manually tracked and corrected tempo[2] and loudness[3] data is available at the beat resolution for multiple recordings of the

[2] `http://mazurka.org.uk/info/excel/beat/`, accessed May 31, 2011.
[3] `http://mazurka.org.uk/info/excel/dyn/gbdyn/`, accessed May 31, 2011.

Mazurka op.	Key	# rec.	Ref. rec.
17 no. 4	A minor	94	Cohen 1997
24 no. 2	C major	65	Meguri 1997
30 no. 2	B minor	60	Sofronitsky 1960
63 no. 3	C# minor	88	Niedzielski 1931
68 no. 3	F major	51	Richter 1976

Table 3.1: Overview of the mazurkas, numbers of different performance recordings used, and the reference recordings used in the experiments described in this chapter.

pieces (for the process, see [Sapp, 2007]).

In this work, we focus on these five mazurkas, thus building on the data acquisition and annotation efforts of the Mazurka Project. For each of the mazurkas, we follow the procedure as outlined in Sections 3.3 and 3.4 and choose the shortest recording for which tracked beat data is available as the reference recording. Choosing the shortest recording reduces the size of the alignment paths, and having manually annotated beat data will be useful for interpreting the results. An overview of the mazurkas used, the number of analyzed audio recordings, and the chosen reference recordings[4] is given in Table 3.1.

3.5.2. EVALUATION STRATEGY

In the work described in this chapter, we focus on qualitative analysis only, with the objective to indicate a relation between the expressive timing deviations between performances and structural and interpretative aspects of a music piece. As stated in Section 3.1, this relation could be exploited to support algorithms for automated analysis, indexing and retrieval of classical music. However, since we are not aiming at formal detection, retrieval or classification at this stage, no formalized or objective ground truth and evaluation methodologies are available in terms of precision or recall. Furthermore, with respect to timing expressivity, while several theories on the reasons for expressive playing have been posed in music psychology and cognition, these have mainly been based on meticulous analyses of detailed performance data obtained for small musical excerpts. It is very hard to transfer these theories to the scope of our full-length mazurkas and the audio domain in such a way that meaningful quantitative evaluation measures can be used. Finally, although the structure of a music piece is largely objective since it is encoded in the score, the indication of exact structure boundaries frequently is subjective and ambiguous, and annotation these precisely in audio recordings is even harder. By using the carefully annotated beat data of the five mazurkas used in our experiment as reference for our structural indications and further discussion, we strive to keep our evaluation as accurate as possible, but even in this case, quantitative evaluation measures are yet to be found.

[4] For more discographic information, see `http://www.mazurka.org.uk/info/discography/`, accessed May 31, 2011.

Mazurka op.	Pearson's corr. coef.
17 no. 4	0.8080
24 no. 2	0.8153
30 no. 2	0.7093
63 no. 3	0.8210
68 no. 3	0.8018

Table 3.2: Pearson's correlation coefficient between standard deviation sequences smoothed with fixed and variable-length windows and sampled at the beat level.

3.6. RESULTS

For all five mazurkas, our methods yielded results that indeed can be interpreted meaningfully in relation to the music content. While we do not have the space to discuss results for each individual mazurka in detail, in this section we will discuss the most striking general results.

3.6.1. SMOOTHING TIMING DEVIATIONS WITH A FIXED MOVING AVERAGE WINDOW

For our experiments, we have chosen a fairly high analysis resolution: with the chosen analysis window length and hop size, in the end we will have a standard deviation value over the full performance ensemble in relation to the reference recording sampled at every 0.0464 s. Therefore, in order to capture informative trends in the timing deviation developments, the standard deviation data needs to be smoothed. This can be done by either applying a fixed-length moving average window, or a variable-length window that depends on higher-level concepts (e.g. smoothing over beat lengths). For each mazurka, we smoothed our short-time standard deviation values in two ways: with a fixed moving average window of length 20 (an empirically found value), and with variable-length windows that corresponded with the manually annotated beat lengths in the reference recording. Subsequently, we sampled the results obtained through the fixed moving average window at the manually annotated beat positions, in order to get one standard deviation value per beat, and compared the values with those obtained using variable-length smoothing. In Table 3.2, Pearson's correlation coefficient between the standard deviation sequences obtained using the two smoothing methods is displayed for each mazurka.

For all mazurkas, the results obtained from both methods are clearly positively correlated. This suggests that applying a fixed moving average window for the data smoothing yields results sufficiently similar to those obtained through variable-length smoothing. This justifies our decision to avoid automated beat-extraction methods (which often introduce inaccuracies themselves) in our approach. With these results, in the remainder of this section, all data values presented will be taken from standard deviation sequences smoothed with a fixed 20-point moving average window.

3.6.2. RELATING TIMING DEVIATIONS TO HIGH-LEVEL MUSICAL STRUCTURE

When applying smoothing using a 20-point moving average window, in each mazurka we can observe periodic patterns of increasing and decreasing standard deviations, that in the next subsection will be shown to relate to musical phrasing. At the same time, when using a larger smoothing window, we can see general increasing and decreasing trends as well, that can be related to main structural boundaries. This can be observed in Figures 3.1 and 3.2, in which timing standard deviations are plotted for two of the mazurkas, smoothed over fixed-length 20- and 100-point windows. In the plots, we have overlaid labels describing the ternary musical structure form of the mazurkas (the upper-case labels), as well as smaller recurring musically thematic blocks (the lower-case labels). These labels were obtained through manual musical structure analysis performed by the authors and have consequently been put at the first strong downbeat of a structural block.

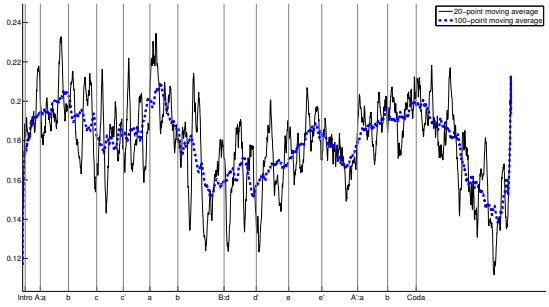


Figure 3.1: Standard deviation trends for mazurka op. 24 no. 2 at different smoothing resolutions.

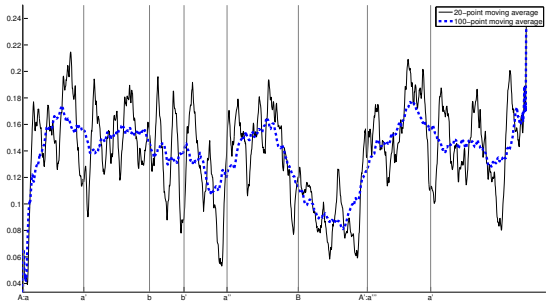


Figure 3.2: Standard deviation trends for mazurka op. 68 no. 3 at different smoothing resolutions.

For the 100-point smoothing windows, we can see general 'trendlines' that change

curvature near structural boundaries, most strongly near the boundaries associated with the main ternary structure. Furthermore, in the case of both mazurka op. 24 no. 2 and op. 68 no. 3, there is a clear difference in average deviation level between the middle 'B' section of the mazurka and the surrounding 'A' sections. In both cases, the 'A' sections show the highest average deviation levels: this would make sense, as, being repeated in a ternary form, they constitute the most characteristic and recurring thematic material of the mazurka and thus will invite a performer to clearly express them.

3.6.3. RELATING TIMING DEVIATIONS TO MUSICAL PHRASING

As stated in the previous subsection (and observable in Figure 3.2), the standard deviation curves show periodic trends, that can be related to musical phrasing structure. In Western classical music, musical phrasing is often built in a hierarchical binary way: an 8-bar phrase can be decomposed in two 4-bar phrases, that typically constitute an antecedent and consequent phrase, which in their turn can often be decomposed into two 2-bar subphrases. Our standard deviation curves strikingly reflect these phrasing patterns. For example, the opening phrase of mazurka op. 68 no. 3 is an 8-bar phrase, containing a (2 + 2)-bar antecedent phrase and a 4-bar consequent phrase. This pattern can be found as well in the corresponding standard deviation curve, shown in Figure 3.3. The curve also shows that in the antecedent phrase, the chords at the first downbeat of both even bars show stronger performer individualism than those opening the odd bars. From a music-theoretical viewpoint, the harmonic developments over these bars from tonic (odd bars) to dominant (even bars) would indeed suggest that the even bars would deserve more expressive emphasis.

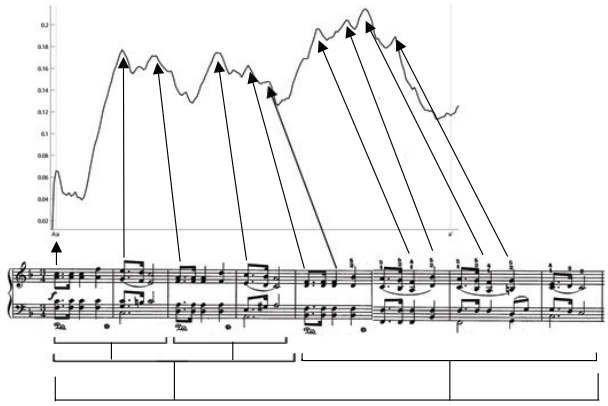


Figure 3.3: Opening phrase of mazurka op. 68 no. 3; curve developments and corresponding score events. Structural phrase hierarchy is indicated below the score.

At several places in the mazurkas, Chopin explicitly indicated an increase of expressiveness in the score. This e.g. is the case in the first phrase of the mazurka op. 17 no. 4 'A' section, where the opening theme is repeated in an ornamented form. The curves reflect the effect of ornamentation on timing individualism, as shown in Figure 3.4: in the large

delicatissimo ornament, a peak building up to the largest deviation among performers is found.

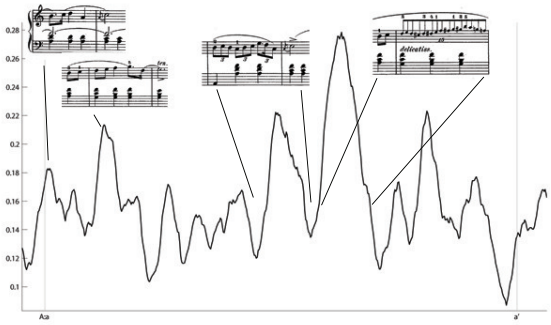


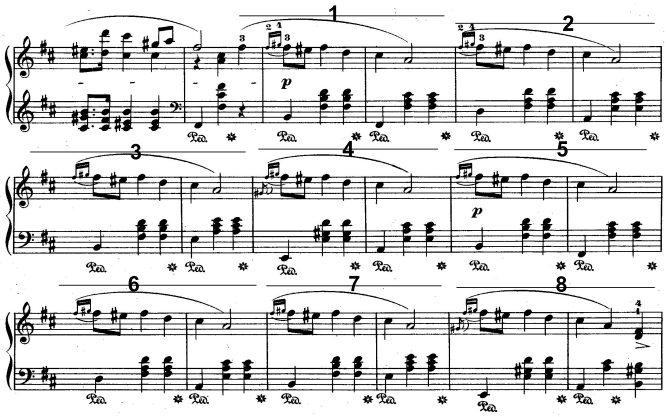
Figure 3.4: Opening phrase of mazurka op. 17 no. 4; curve developments related to ornamentation.

However, the curves also reveal information that cannot directly be inferred from the score. For example, the 'c' section of mazurka op. 30 no. 2 (Figure 3.5(a)) consists of eight nearly identical repeats of the same two-bar subphrase, which can hierarchically be grouped as two 'sentences' of four subphrases, which each have an antecedent and consequent phrase encompassing two subphrases. One would expect that there will be expressive deviation differences between these subphrase statements. Indeed, one can see in Figure 3.5(b) that such differences are present and related to the phrase structure. The largest deviations occur when the second four-subphrase sentence is started; furthermore, we notice that in the first sentence, the start of the consequent phrase has a lot of expressive deviation, while in the second sentence, there is a more gradual build-up towards the end.

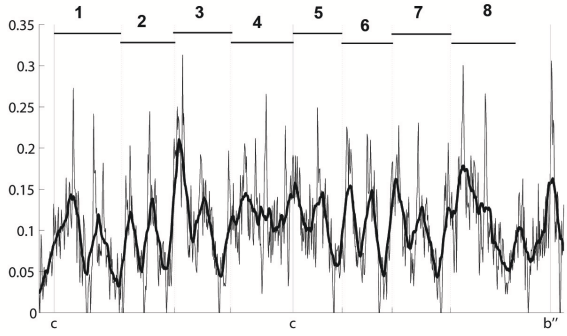
3.6.4. ROBUSTNESS TO NOISE AND REFERENCE RECORDING

At the moment, our method does not take into account any signal noise filtering, which can be risky, especially when analyzing historical recordings. Indeed, at the very end of the curves shown in Figures 3.1 and 3.2, a sudden standard deviation increase is observed, caused by irregular alignments between (noisy) silences in recordings occurring after the last musical notes are played. Furthermore, for mazurka op. 63 no. 3, where our reference recording is a digital transfer of a 78 rpm recording from 1931, the largest standard deviation value found for the piece corresponds with a hiss in the recording, rather than a musical event.

Nevertheless, when sampling the standard deviation values at the manually annotated beat locations in the 1931 Niedzielski recording, and comparing the resulting curve to beat-sampled standard deviation values obtained from other, cleanly recorded reference recordings, the Niedzielski recording does not appear to yield a much noisier curve than the clean recordings, as shown in Figure 3.6. However, it is striking to see that the curve developments for different recordings do not always follow the same development trend, although several peaks stand out consistently for all recordings. This suggests that



(a) Score with numbered subphrases



(b) Standard deviation curve with numbered subphrases

Figure 3.5: Mazurka op. 30 no. 2, 'c' section; curve developments related to identical phrase repeating.

the timing standard deviation curve is partially biased towards the reference recording, and that considering multiple reference recordings will yield a more objective view on the expressive timing opportunities within a piece. This needs further investigation in our future work.

3.7. CONCLUSION AND RECOMMENDATIONS

By focusing on timing deviations between multiple recorded performances of the same musical piece, our audio alignment analysis methods are capable of highlighting timing expressivity opportunities in Chopin Mazurkas which can be interpreted in musically meaningful ways. Smoothing the obtained standard deviation values with fixed-length moving average windows yields similar results to those obtained through variable-length smoothing over the beats, while being considerably simpler and less sensitive to intermediate processing errors. The developments found in the standard deviation curves

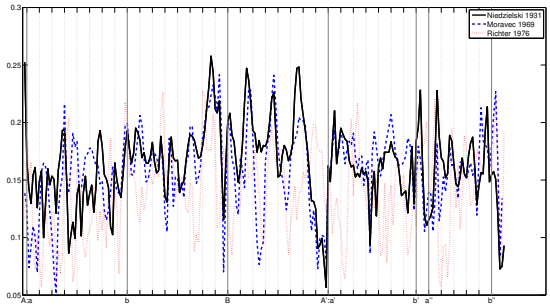


Figure 3.6: Mazurka op. 63 no. 3, standard deviations sampled at beat markings for Niedzielski, Moravec and Richter recordings.

turn out to relate very well to music-theoretical concepts of overall musical structure and phrasing. In addition, the standard deviation curves also indicate expression opportunities that are not obvious from an objective reading of a score. Finally, the method appear to be relatively noise-robust, although true noise-suppressing methods still need to be incorporated in our approach.

From comparisons between different reference recordings, it appeared that the results of our method may potentially be biased towards the chosen reference recording, but this is a topic that should be investigated further in our future work. Other useful directions for further work include the application of our method to data genres other than Chopin mazurkas, and to specific performer data sets. For example, it will be interesting to investigate whether the moments in a piece that show the largest expression variance among performers are invariant to different performer schools or traditions (e.g. historical recordings vs. modern recordings, Russian school vs. western schools). It also will be worthwhile to research whether other expressivity parameters can be extracted from the audio data that can be used to enhance our results. Finally, while it still was infeasible for this chapter, it will be interesting to see if the psychological and cognitive models on music expression that have been proposed earlier (e.g. on tension-relaxation patterns) can be transferred to music audio evaluation settings. This will help in making evaluation results further generalizable.

By any means, it is promising that our low-level approach can extract part of the domain knowledge and artistic understanding of musical performers, that implicitly is encoded in the audio signal. Although our method does not provide a concrete detection or identification mechanism yet, as stated in our Introduction, it will be capable of providing supporting information for multiple Music-IR tasks, and as such can aid in the high-level interpretation of results from these tasks.

4

EXTENDING THE CASE STUDY ON TIMING DEVIATIONS

As discussed in the previous chapter, audio recordings of classical music pieces reflect the artistic interpretation of the piece as seen by the recorded performing musician. With many recordings being typically available for the same music piece, multiple expressive rendition variations of this piece are obtained, many of which are induced by the underlying musical content. In the previous chapter, we focused on timing as a means of expressivity, and proposed a light-weight, unsupervised and audio-based method to study timing deviations among different performances through alignment patterns. By using the standard deviation of alignment patterns as a measure for the display of individuality in a recording, structural and interpretational aspects of a music piece turned out to be highlighted in a qualitative case study on five Chopin mazurkas. In this chapter, we propose an entropy-based deviation measure as an alternative to the existing standard deviation measure. The obtained results for multiple short-time window resolutions, both from a quantitative and qualitative perspective, strengthen our earlier finding that the found patterns are musically informative and confirm that entropy is a good alternative measure for highlighting expressive timing deviations in recordings.

The contents of this chapter previously were published as Cynthia C. S. Liem and Alan Hanjalic. Expressive Timing from Cross-Performance and Audio-based Alignment Patterns: An Extended Case Study. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 519–524, Miami, Florida, USA, October 2011.

4.1. INTRODUCTION

In classical music, music pieces are usually conceived by composers and translated into scores. These are studied and interpreted by musicians, who each give their own personal, expressive account of the score through their actual performance of the piece. With an increasing number of such performances becoming available in digital form, we also gain access to many different artistic readings of music pieces.

The availability of recordings of multiple performances of music pieces previously has strongly been exploited in the field of audio similarity-based retrieval. In this, the focus was on matching musically closely related fragments (*audio matching* [Müller, 2007, Müller et al., 2005]), or finding different versions of a song at the document level, ranging from different performances of the same notated score (*opus retrieval* [Casey et al., 2008b]) to potentially radically different new renditions of a previously recorded song (*cover song identification* [Serrà et al., 2008]). In general, matching and retrieval of classical music pieces were shown to be achievable with near-perfect results [Casey et al., 2008a, Liem and Hanjalic, 2009]. Another category of previous work largely focused on analyzing and/or visualizing the playing characteristics of individual performers in comparison to other performers [Grachten and Widmer, 2009, Sapp, 2007, 2008].

At certain moments, a performer will display larger personal expressive freedom than at other moments, guided by theoretical and stylistic musical domain knowledge as well as personal taste and emotion. By comparing expressive manifestations in multiple recordings of the same piece, we therefore can gain insight in places in the piece where the notated musical content invites performers to display more or less expressive individualism. Such information on the interplay between performance aspects and the notated musical content provides a novel perspective on the implicit interpretative aspects of the content, which can be of a direct benefit for many Music Information Retrieval (Music-IR) tasks, ranging from music-historical performance school analysis to quick and informed differentiating and previewing of multiple recordings of the same piece in large databases.

In recent previous work [Liem et al., 2011a], we proposed a light-weight, unsupervised and audio-based method to study timing deviations among different performances. The results of a qualitative study obtained for 5 Chopin mazurkas showed that timing individualism as inferred by our method can be related to the structure of a music piece, and even highlight interpretational aspects of a piece that are not necessarily visible from the musical score. In this chapter, we introduce an entropy-based approach as an alternative to our previous standard deviation-based approach, and will study the characteristics of both methods in more depth at multiple short-time window resolutions. While this task does not have a clear-cut ground truth, the introduction of our new entropy method allows for quantitative comparative analyses, providing deeper and more generalizable insight into our methods than the largely qualitative pioneering analyses from [Liem et al., 2011a].

This chapter is organized as follows. After a summary of our previous work from [Liem et al., 2011a], we will describe our new entropy-based method. This will be followed by a description of the experimental setup and corresponding results. Finally, the chapter will end with a conclusion and discussion of future directions.

4.2. AUDIO-BASED ALIGNMENT AND ANALYSIS OF MULTIPLE PERFORMANCES

4.2.1. AUDIO-BASED ALIGNMENT OF MULTIPLE PERFORMANCES

In [Liem et al., 2011a], we proposed a method to infer timing expressivity in an audio-based, objective and unsupervised data-driven way, largely building on novel work in audio similarity-based retrieval.

As short-time harmonic audio signal descriptor features, we adopt the recent Chroma Discrete Cosine Transform-reduced Log Pitch (CRP) features, which outperformed traditional chroma representations in timbre-robustness and audio matching performance [Müller and Ewert, 2010]. We use the CRP feature implementation as made available by the original authors[1]. If A is a set with n audio recordings of the same piece, we obtain n CRP profile vectors r establishing a set R, where each r represents an audio recording $a \in A$.

As different performances of the same piece may differ in global tempo, the CRP profile vectors $r \in R$ will have different lengths. Through Dynamic Time Warping (DTW) techniques, we can align the vectors and find a time mapping between corresponding events in different recordings. For this, we apply the DTW alignment technique from [Serrà et al., 2008], which used a binary cost measure and imposed local constraints to avoid pathological warpings. This method was shown to be very powerful in cover song retrieval settings. We choose a CRP profile vector $r_{ref} \in R$, corresponding to a reference recording that may be arbitrary chosen. By aligning r_{ref} with the vectors $r \in R \backslash \{r_{ref}\}$, corresponding to all other recordings in the set, full alignment between performances is achieved through r_{ref}. For each alignment between r_{ref} and an $r \in R$, an alignment matrix X is constructed. The alignment value $X_{i,j}$ between two CRP profiles at time instances i and j in r_{ref} and r, respectively ($r_{ref}[i]$ and $r[j]$), is computed adopting the local constraints as suggested in [Serrà et al., 2008]. Initialization procedures, binary similarity measures and other parameters were also taken from this article, to which the interested reader is referred for more details.

An explicit alignment path is obtained by tracing back from the point corresponding to the highest total alignment score. If $|r_{ref}| = m$, for each alignment to a performance r we obtain an alignment path w of length m, with $w[1 \dots m]$ indicating short-time instance indices of the CRP profiles in r that align to $r_{ref}[1 \dots m]$. Not all time instances $1 \dots m$ may have been explicitely covered in the original alignment path. Assuming linear development for unknown instances, missing values are estimated through linear interpolation.

4.2.2. PERFORMANCE ALIGNMENT ANALYSIS

After calculating all alignment paths following the procedures above, we will have obtained a set W with $n-1$ alignment paths $w \in W$, each of length m. We post-process these paths to emphasize irregular alignment behavior: if an alignment subpath $w[k \dots l]$ shows constant alignment steps ($w[k] = w[k+1] = w[k+2] = \dots = w[l-1] = w[l]$), this means that the corresponding CRP feature vector excerpt in r is a linearly scaled version of $r_{ref}[k \dots l]$, and therefore does not reflect any timing individualism. In order to

[1] `http://www.mpi-inf.mpg.de/~mmueller/chromatoolbox/`, accessed August 18, 2011.

highlight alignment step slope changes, we compute discrete second derivatives over the alignment path.

First of all, for each alignment path w, we compute the discrete first derivative δ through the central difference:

$$\delta[i] = \begin{cases} \frac{1}{2}(w[i+1] - w[i-1]) & 1 \leq i \leq m \\ w[1] - w[0] & i = 1 \\ w[m] - w[m-1] & i = m. \end{cases}$$

Due to an initial alignment index jump, a large 'startup' derivative is found at the beginning of the path. As we are only interested in the alignment step development within the true alignment path (and the beginning of the recording for the given time sampling rate will contain silence), we set the derivative values up to this startup point to 0. By repeating the central difference procedure on the enhanced δ, a second derivative approximation $\delta^2 \in \Delta^2$ is obtained.

We assume that moments in the piece showing the largest timing deviations among performers (and thus, the highest degree of individualism) must have given the performers a reason to do so, and therefore must be of a certain semantic relevance. A measure is needed to express this individuality of timing at all short-time instances of Δ^2. For this, we proposed to adopt the standard deviation: for each time instance $t = 1 \ldots m$, we compute $\sigma[t]$, which is the standard deviation of all alignment second derivatives $\delta^2[t] \in \Delta^2$, acquiring a standard deviation sequence σ of length m.

4.3. ENTROPY AS INFORMATION MEASURE

The assumption that moments with the largest timing deviations ('disagreement') among performers will be of a certain semantic relevance resembles the notion of entropy in information theory, where items with the most uncertain actual realization are considered to hold the largest amount of information. Thus, as an alternative to our previous standard deviation method, we now propose to calculate the entropy of Δ^2 at each short-time instance. If Δ^2 has the possible values ('symbols') $d^2_{t,1} \ldots d^2_{t,f}$ at time t, then

$$h[t] = -\sum_{i=1}^{f} p(d^2_{t,i}) \log_2 p(d^2_{t,i})$$

where we approximate $p(d^2_{t,i})$ by the frequency of $d^2_{t,i}$ in Δ^2 at time instance t. While the previous standard deviation-based approach treats the values at each $\delta^2[t]$ as cardinal data, the entropy-based approach will treat the values as nominal data, only measuring diversity.

4.4. EXPERIMENTAL EVALUATION

We initially conceived our methods with the goal to reveal implicitly encoded expressive musical information in audio that would go beyond an objective score reading. This means that no explicit classification is applied and an objective ground truth is absent. Because of this, in [Liem et al., 2011a], the results of the standard deviation-based method were largely discussed in a qualititative way. With our new entropy-based

method, possibilities arise for quantitative comparisons between this method and the standard deviation-based method, which we will discuss in this section, as an addition to qualitative and musical interpretations of the results of the entropy-based method.

Our experiments will focus on two aspects: (1) verifying that σ and h are no random noise sequences and (2) focusing on the main similarities and dissimilarities between σ and h from a quantitative and qualitative perspective. While the work in [Liem et al., 2011a] only focused on a 2048-sample short-time audio analysis window, our current experiments will consider multiple possible window lengths. While we are not striving to identify an 'optimal' time window length yet (which will depend on the desired musical unit resolution, e.g. small ornamental notes vs. harmonies on beats), we consider these multiple window lengths to verify if the behavior of our methods is stable enough to not only yield interpretable results at the earlier studied resolution of 2048 samples.

4.4.1. EXPERIMENTAL SETUP

Following our earlier work, we focus on 5 Chopin mazurkas that were thoroughly annotated as part of the CHARM Mazurka Project [Sapp, 2007]: op. 17 no. 4, op. 24 no. 2, op. 30 no. 2, op. 63 no. 3 and op. 68 no. 3, with 94, 65, 60, 88 and 51 available recordings, respectively. We follow the procedure as outlined in Section 4.2.1, choosing the shortest recording for which manually annotated beat data is available as the reference recording, thus minimizing the size of the alignment paths. In order to interpret the results, we will use manual musical structure analyses by the authors as a reference. Thanks to the carefully established manual beat annotations from the Mazurka dataset, these structure analyses can be related to the audio as precisely as possible.

We apply our methods to all available recordings in each of the mazurkas, calculating standard deviations σ and entropies h for the alignment pattern second derivatives in Δ^2, as obtained for 7 different short-time window lengths (from 1024 to 4096 samples, in linearly increasing steps of 512 samples, at a sampling frequency of 22050 Hz and with 50% overlap). A representative example of second derivative value frequencies over the short-time instances is shown in Figure 4.1: the majority of values is zero ('constant alignment development'), and frequency peaks for other values appear to occur in bursts.

4.4.2. VERIFICATION OF TRENDS IN STANDARD DEVIATIONS AND ENTROPIES

To verify that both the sequences σ and h are no random noise sequences, we perform two statistical runs tests: one testing the distribution of values above and under the sequence mean, and one testing the distribution of upward and downward runs. In both cases and for all window lengths, the tests very strongly reject the null hypothesis that the sequences are random. In Figure 4.2, the runs frequencies for the test focusing on upward and downward runs are plotted. From this plot, we notice that entropy sequences consistently have less up- and downward runs (and thus 'smoother behavior') than standard deviation sequences, especially for small window sizes. Furthermore, the relation between the number of runs and the window size does not appear to be linear, implying that the choice of a larger short-time window does not uniformly smooth the results obtained with a smaller window. Curves for the test focusing on values above and un-

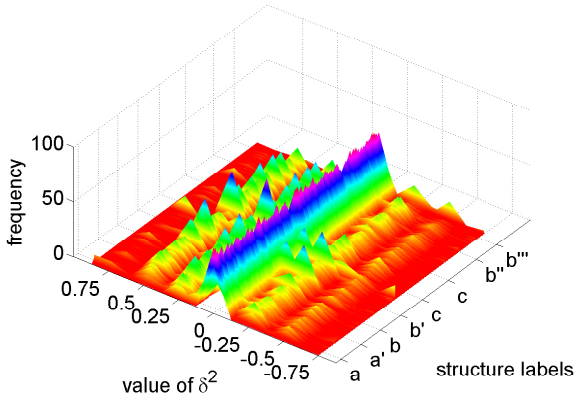


Figure 4.1: Histogram for δ^2 values in Δ^2 measured at consecutive short-time windows for mazurka op. 30 no. 2, for a 2048-sample window length and with reference main structural boundary labels (a, b, c, etc.) indicated over the time dimension.

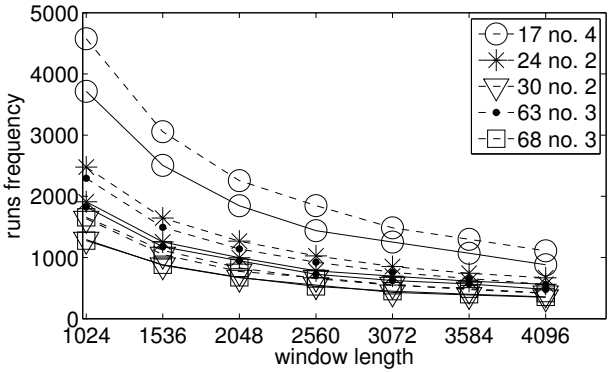


Figure 4.2: Numbers of up- and downward runs (summed) for different short-time window lengths. Dashed lines indicate σ sequences, solid lines indicate h sequences. Markers indicate mazurkas.

der the sequence mean are omitted due to space considerations, but strongly resemble the given plot. When plotting the resulting sequences over time, the resulting h curves indeed are less noisy than the σ curves. Figure 4.3 shows both curves for the opening phrase of mazurka op. 17 no. 4 for a short-time window of 1024 samples. The σ curve appears to be denser, due to the larger number of up- and downward runs. Looking at the general development of the curves, both σ and h appear to show very similar behavior, with many co-occurring maxima and minima. As a quantitative backing for this notion, Table 4.1 shows Pearson's correlation coefficient between σ and h for all window lengths considered. From the values in this table, it indeed becomes clear that σ and h are strongly correlated.

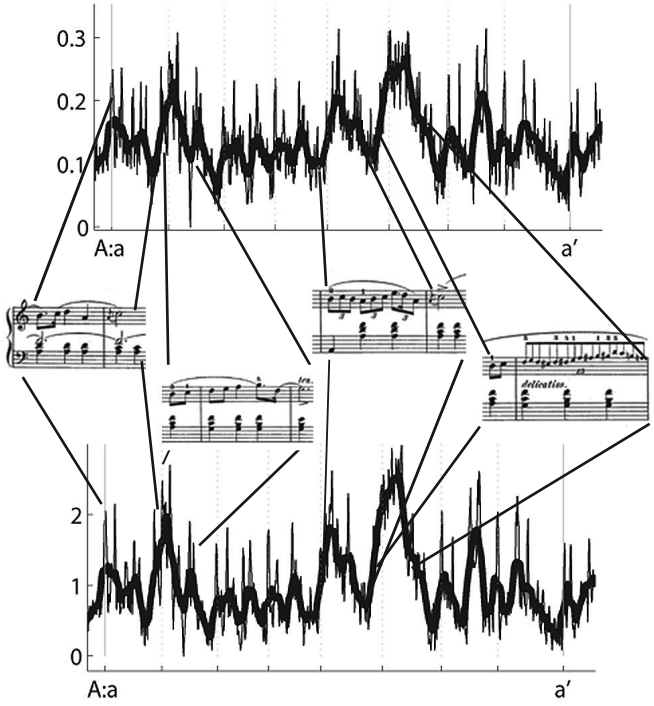


Figure 4.3: σ (top) and h (bottom) sequence for opening phrase of mazurka op. 17 no. 4 with corresponding score fragments. 1024-sample window length, 20-point moving average smoothed trendline indicated with thick line.

	1024	1536	2048	2560	3072	3584	4096
17 no. 4	0.9271	0.9225	0.9184	0.9117	0.9089	0.9022	0.9007
24 no. 2	0.9352	0.9308	0.9245	0.9218	0.9104	0.9105	0.9045
30 no. 2	0.9107	0.9094	0.9138	0.8955	0.8952	0.8911	0.8945
63 no. 3	0.9165	0.9103	0.9113	0.8992	0.8930	0.8877	0.8876
68 no. 3	0.9261	0.9274	0.9302	0.9387	0.9333	0.9291	0.9321

Table 4.1: Pearson's correlation coefficient between σ and h sequences for all five mazurkas with different short-time window lengths (in samples).

4.4.3. STANDARD DEVIATIONS VS. ENTROPIES

As mentioned above, entropy sequences h are strongly correlated with standard deviation sequences σ. Thus, as with the σ sequences, they will be capable of highlighting developments that musically make sense [Liem et al., 2011a]. Next to the example in Figure 4.3, where both the σ and h values increased with ornamentational variation, we

	1024	1536	2048	2560	3072	3584	4096
17 no. 4 overall	0.2736	0.2595	0.3994	0.3413	0.4303	0.2847	0.6966
17 no. 4 at beat starts	0.4217	0.3460	0.4798	0.3662	0.4571	0.2955	0.7020
17 no. 4 at subphrase starts	0.6462	0.5077	0.6769	0.4769	0.5231	0.4462	0.7385
24 no. 2 overall	0.3645	0.5912	0.3172	0.4754	0.6417	0.5548	0.7307
24 no. 2 at beat starts	0.4903	0.6842	0.3767	0.5097	0.6898	0.5845	0.7895
24 no. 2 at subphrase starts	0.5085	0.7288	0.3559	0.5254	0.7966	0.6271	0.8644
30 no. 2 overall	0.2238	0.2354	0.1944	0.1790	0.3030	0.4177	0.6508
30 no. 2 at beat starts	0.3212	0.3005	0.1606	0.1762	0.2902	0.4301	0.6321
30 no. 2 at subphrase starts	0.4375	0.4375	0.3125	0.3438	0.3750	0.5000	0.8125
63 no. 3 overall	0.4901	0.5869	0.7861	0.6578	0.8038	0.5617	0.5956
63 no. 3 at beat starts	0.6348	0.6565	0.8348	0.6696	0.8261	0.5435	0.5739
63 no. 3 at subphrase starts	0.8684	0.8947	0.9474	0.7895	0.8421	0.5789	0.6053
68 no. 3 overall	0.1574	0.3359	0.1383	0.2698	0.6095	0.4751	0.6628
68 no. 3 at beat starts	0.3039	0.4420	0.1823	0.3094	0.6575	0.5304	0.6906
68 no. 3 at subphrase starts	0.3000	0.5000	0.2333	0.4000	0.6333	0.7000	0.7000

Table 4.2: Normalized entropies h_{norm} vs. standard deviations σ_{norm}: fractions of cases in which $h_{norm} > \sigma_{norm}$ considered over all short-time instances, over all beat starts, and over all subphrase starts different short-time window lengths (in samples).

also give an example where the musical score does not clearly indicate the expressive development of phrases. In Figure 4.4, the 'c' section of mazurka op. 30 no. 2 is shown, where a simple subphrase is almost identically repeated 8 times. A performer will not play this subphrase 8 times in an identical way, and this is reflected both in σ and h: the major displays of individuality in recordings can be found in subphrases 1 (first statement of subphrase), 3 (following traditional binary period structures, here a new subphrase could be starting, but this is not the case) and 8 (last statement of subphrase). Furthermore, for subphrase 4 and 8, the average value of σ and h is higher than in the other subphrases, and no minima are reached as large as in the other phrases. This can be explained because of the altered ornament starting the subphrase, and the fact that both subphrase 4 and 8 are the final subphrase in a higher-order phrase hierarchy of 4 + 4 subphrases. From both Figure 4.3 and 4.4, the main difference between σ and h appears to be that h has a considerably larger range than σ, and especially tends to amplify positive peaks.

With its less noisy behavior and stronger peak amplification, the entropy-based method seems more attractive for our alignment analyses than the standard deviation-based method. As a final experiment aimed at gaining more insight into the differences between both methods, we linearly scale both σ and h to unit range. This results in sequences σ_{norm} and h_{norm}. We then test how often $h_{norm} > \sigma_{norm}$ for three cases: (1) all short-time instances, (2) all beat starts (with the beat timings obtained from the earlier manual annotations from the CHARM project) and (3) all subphrase starts. While these cases consider a decreasing number of events, the musical importance of the events increases: a subphrase start should be more informative than a random instance in time. Results are given in Table 4.2.

In general, σ_{norm} will have larger values than h_{norm}. This matches with the notion that the entropy sequences amplify positive peaks: thus, the non-peak values will tend to skew under the mean entropy value, while standard deviations are centered around the

mean in a more balanced way. Mazurka op. 63 no. 3 is an exception, but this may have been caused by the noisiness of the historical reference recording (Niedzielski 1931), which causes clicking and hissing effects at random moments throughout the piece, thus also causing irregular alignment behavior at these random moments. However, in all cases, when only looking at time instances with beat and subphrase starts, the fraction of larger normalized entropies increases for all mazurkas. Especially for subphrases in comparison to beat starts, the increase is considerable. This implies that the entropy sequence values indeed amplify musically meaningful peaks.

Looking at the differences between beat start and subphrase start fractions, the increases initially may not appear to be stable or generalizable over different mazurkas. For subphrase starts, the probability that $h_{norm} > \sigma_{norm}$ is much larger than for beat starts in mazurkas op. 17 no. 4 and op. 63 no. 3 (and to a lesser extent, op. 30 no. 2). On the other hand, in mazurkas op. 24 no. 2 and op. 68 no. 3, this is much less the case, with the beat and subphrase start fractions being much closer to each other.

From a musical perspective, this may not seem as strange as from a numerical perspective: mazurkas op. 24 no. 2 and op. 68 no. 3 both are rather 'straightforward' pieces, with many repeating blocks with little thematic development, and constant ongoing rhythms. Thus, there is not so much flexibility to shape structural boundaries and subphrase starts with large timing differences. On the other hand, mazurkas op. 17 no. 4 and op. 63 no. 3 are very dramatical, have strongly differing thematic blocks, and thus allow for emphasizing of new subphrases. While resembling mazurkas op. 24 no. 2 and op. 68 no. 3 in terms of rhythmical and thematic straightforwardness, mazurka op. 30 no. 2 is less rigid in terms of phrasing and musical movement, and thus will allow for more timing flexibility, thus also sharing characteristics with the other two mazurkas.

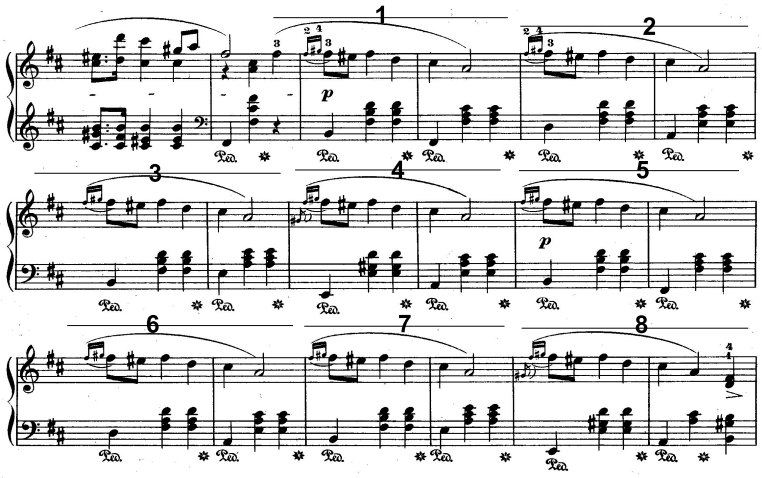
4.5. CONCLUSION AND RECOMMENDATIONS

In this chapter, we proposed an entropy-based method as an alternative to a standard deviation-based method for studying alignment patterns between multiple audio recordings, which were considered to contain interesting information about the recorded music that cannot objectively be inferred from a score. Our entropy method yielded results that consistently were strongly correlated with the standard deviation results at multiple time resolutions, while being less noisy and amplifying positive peaks, which both are desirable properties for our purposes. It was shown that both the standard deviation and entropy methods do not depict random noise, but can be related to actual musical content.

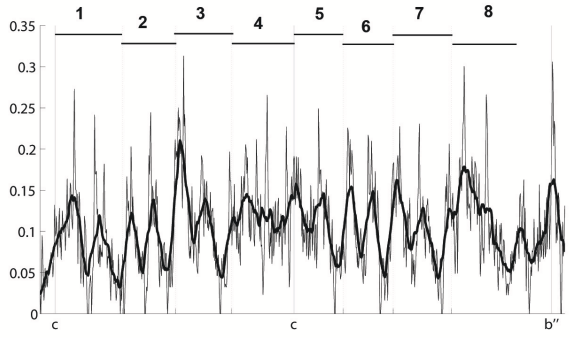
The development over multiple time resolutions of correlations between standard deviation and entropy sequences, the frequencies of up- and downward runs, as well as runs above and under the sequence mean, yields similar trends over different mazurkas, implying that our methods are generalizable. We did not focus yet on further implications of the choice of short-time window length, which still needs to be done in future work. Another main future challenge is the further solidification and backing of the musical interpretations of our results. Finally, we did not yet employ any noise-filtering or signal enhancement techniques. While the results obtained for the noisy op. 68 no. 3 Niedzielski reference recording on runs frequency and correlation trends are largely consistent with the results for other mazurkas with clean reference recordings, the refer-

ence recording quality will influence results and this topic should be investigated more in future work.

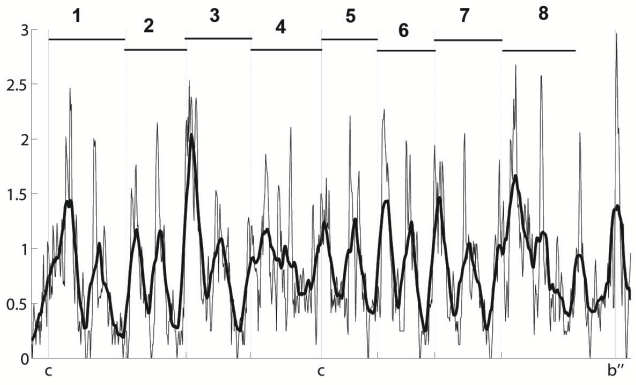
Rendering MIDI files as audio and modifying them in a controlled way may partially overcome the problem of a missing ground truth and possible noise in real-life reference recordings. In addition, the interpretation of results can be strengthened through a combination of our methods with other Music-IR techniques dealing with prior knowledge of the musical content in a more explicit and supervised way. Supported by our methods, such techniques will not have to be tediously applied to a full database, but can be limited to one or more reference recordings. This introduces promising directions for Music-IR tasks dealing with the real-life abundance of artistically valuable digital recordings.



(a) Score with numbered subphrases



(b) Standard deviation sequence σ



(c) Entropy sequence h

Figure 4.4: Mazurka op. 30 no. 2, σ and h for 'c' section. The 8 repeating subphrases are numbered. 1024-sample window length, 20-point moving average smoothed trendline.

5

COMPARATIVE ANALYSIS OF ORCHESTRAL PERFORMANCE RECORDINGS: AN IMAGE-BASED APPROACH

Traditionally, the computer-assisted comparison of multiple performances of the same piece focused on performances on single instruments. Due to data availability, there also has been a strong bias towards analyzing piano performances, in which local timing, dynamics and articulation are important expressive performance features. In this chapter, we consider the problem of analyzing multiple performances of the same symphonic piece, performed by different orchestras and different conductors. While differences between interpretations in this genre may include commonly studied features on timing, dynamics and articulation, the timbre of the orchestra and choices of balance within the ensemble are other important aspects distinguishing different orchestral interpretations from one another. While it is hard to model these higher-level aspects as explicit audio features, they can usually be noted visually in spectrogram plots. We therefore propose a method to compare orchestra performances by examining visual spectrogram characteristics. Inspired by eigenfaces in human face recognition, we apply Principal Components Analysis on synchronized performance fragments to localize areas of cross-performance variation in time and frequency. We discuss how this information can be used to examine performer differences, and how beyond pairwise comparison, relative differences can be studied between multiple performances in a corpus at once.

The contents of this chapter previously were published as Cynthia C. S. Liem and Alan Hanjalic. Comparative Analysis of Orchestral Performance Recordings: an Image-Based Approach. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 302–308, Málaga, Spain, October 2015.

5.1. INTRODUCTION

A written notation is not the final, ultimate representation of music. As Babbitt proposed, music can be represented in the acoustic (physical), auditory (perceived) and graphemic (notated) domain, and as Wiggins noted, in each of these, projections are observed of the abstract and intangible concept of 'music' [Wiggins, 2009]. In classical music, composers usually write down a notated score. Subsequently, in performance, multiple different musicians will present their own artistic reading and interpretation of this score.

Nowadays, increasing amounts of digital music recordings become available. As a consequence, for musical pieces, an increasing amount of (different) recorded performances can be found. Therefore, in terms of data availability, increasing opportunities emerge to study and compare different recordings of the same piece. Beyond the Music Information Retrieval (Music-IR) domain, this can serve long-term interests in psychology and cognition on processes and manifestations of expressive playing (e.g. [Desain and Honing, 1994, Penel and Drake, 1998, Seashore, 1938]), while the analysis of performance styles and schools also is of interest to musicologists [Cook, 2005, Liebman et al., 2012].

In this chapter, we mostly are interested in the analysis of multiple performances of the same piece from a search engine and archive exploration perspective. If one is looking for a piece and is confronted with multiple alternative performances, how can technology assist in giving overviews of main differences between available performances? Given a corpus, are certain performances very similar or dissimilar to one another?

In contrast to common approaches in automated analysis of multiple performances, we will not depart from explicit modeling of performance parameters from a signal. Instead, we take a more holistic approach, proposing to consider spectrogram images. This choice has two reasons: first of all, we are particularly interested in finding methods for comparative analysis of orchestra recordings. We conjecture that the richness of orchestra sounds is better captured in spectrogram images than in mid-level audio features. Secondly, as we will demonstrate in this chapter, we believe spectrogram images offer interpretable insights into performance nuances.

After discussing the state-of-the-art in performance analysis in Section 5.2, in Section 5.3, we will further motivate our choice to compare performances through visual comparison of spectrogram images. Subsequently, Section 5.4 details our chosen comparison method, after which we present the experimental setup for this chapter in Section 5.5. We will then illustrate our approach and its outcomes through a case study in Section 5.6, with a detailed discussion of selected musically meaningful examples. This is followed by a discussion on how our method can assist corpus-wide clustering of performances in Section 5.7, after which the Conclusion will be presented.

5.2. STATE-OF-THE-ART REVIEW

A lot of work exists on analyzing musical performance expressivity. In several cases, establishing models for computer-rendered expressive performances was the ultimate goal (e.g. see [Goebl and Widmer, 2008, Goebl et al., 2007]). Other works focused on identifying reasons behind performance expressivity, including lower-level perceptual

processes [Penel and Drake, 1998]; varying score editions, individual treatments of ornamentation and pedaling, and music-theoretic notions of expectation and tension-relaxation [Palmer, 1996]; generative rules, emotional expression, random variability, motion principles and stylistic unexpectedness [Juslin, 2003]; and musical structure [Friberg and Sundberg, 1999, Grachten and Widmer, 2009, Palmer, 1996]. Historically, the analysis of musical performance strongly focused on expressivity in piano playing (e.g. [Desain and Honing, 1994, Palmer, 1996, Penel and Drake, 1998, Repp, 1994]). The few exceptions to this rule focused on violin performance (e.g. [Cheng and Chew, 2008]), movement in clarinet players (e.g. [Teixeira et al., 2014]), and performance of trained and untrained singers (e.g. [Devaney et al., 2011], inspired by [Seashore, 1938]), but to the best of our knowledge, no systematic comparative studies have been performed considering larger ensembles.

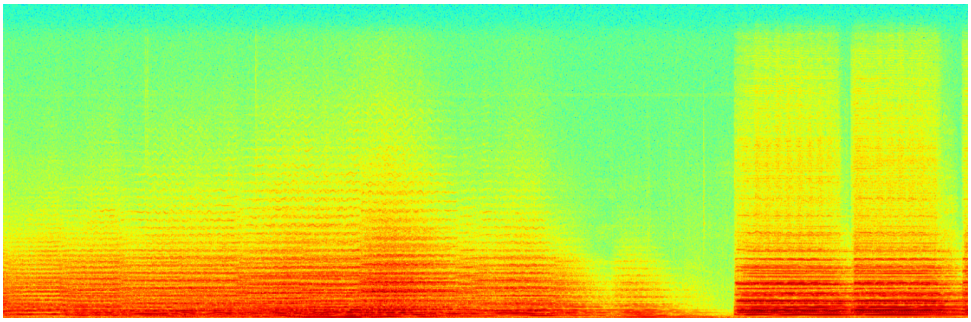
A reason for the general bias towards piano performance may be that digital player pianos (e.g. the Yamaha Disklavier) allow a very precise recording of mechanical performance parameters. When such parameters are available, inter-onset-intervals (IOIs), expressing the time between subsequent onsets, are frequently studied. Otherwise, performance parameters have to be extracted or annotated from the audio signal. As a piano has a discrete pitch set and percussive mechanics, expressive possibilities for a pianist are restricted to timing, dynamics and articulation. As a consequence, audio-based performance analysis methods usually focus on local timing and dynamics. Since it is not trivial to find a suitable time unit for which these parameters should be extracted, supervised or semi-supervised methods often have been applied to obtain this, e.g. by departing from manually annotating beat labels (e.g. [Sapp, 2007, 2008]). However, it is hard (if not infeasible) to realize such a (semi-)supervised approach at scale. Therefore, while a very large corpus of recorded Chopin Mazurkas exists, in practice only the Mazurkas for which annotated beat information exists have been studied in further depth (e.g. [Kosta et al., 2014, Müller et al., 2010b, Sapp, 2007, 2008]).

Alternatively, in [Liem and Hanjalic, 2011, Liem et al., 2011a] an unsupervised approach for comparing Mazurka recordings was proposed which does not rely on explicitly modeled higher-level performance parameters or semantic temporal units, but rather on alignment patterns from low-level short-time frame analyses. As such, this approach would be scalable to a larger corpus. Furthermore, while the choice of not adopting explicit performance parameters makes evaluation of a clear-cut ground truth less trivial, at the same time it allows for any salient variations to emerge automatically from the analysis. The work of this chapter follows a similar philosophy.

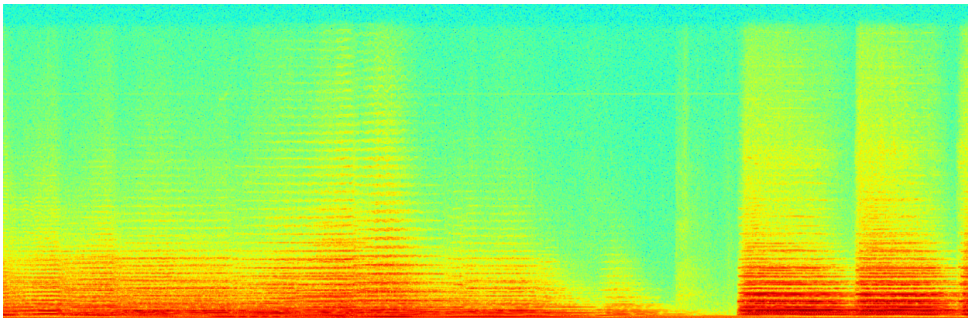
5.3. MOTIVATION FOR SPECTROGRAM IMAGES

In this chapter, we focus on the comparative analysis of orchestra recordings. An orchestra involves a mix of many instruments. Hence, the overall orchestral sound is richer than that of a piano, although individual beat placings and note onsets will be much smoother. Given the multitude of involved players, an orchestra needs guidance by a conductor. Due to this coordinated setup, there is less room for individual freedom in both local dynamics and tempo than in Romantic piano music repertoire. Thus, while local tempo deviations still occur in orchestral recordings, one cannot expect these to reflect performer individuality as strongly as for example in the case of Chopin Mazurkas.

At the same time, in terms of timbre, balance and phrasing articulation, a conductor has a much richer palette than isolated instruments can offer. These aspects are not trivial to explicitly model or interpret from audio signals. However, relevant information may be reflected in recording spectrograms, as illustrated in Figure 5.1. While it is hard to point out individual instruments, a spectrogram can visually reveal how rich the overall sound is, where signal energy is concentrated, and if there are any salient sound quality developments over time, such as vibrato notes.



(a) Georg Solti, Chicago Symphony Orchestra, 1973.



(b) Nikolaus Harnoncourt, Chamber Orchestra of Europe, 1990.

Figure 5.1: Beethoven's Eroica symphony, 2nd movement, spectrogram of bars 56-60 for two different interpretations.

Indeed, spectrograms are commonly used in audio editing tools for visualization, navigation and analysis purposes. In an ethnographic study of musicologists studying historical recordings, it further was shown that examination of the spectrogram helped musicologists in discovering and listening to performance nuances [Barthet and Dixon, 2011]. Therefore, regarding potential end users of performance analysis and exploration tools, spectrogram images may be more familiar and interpretable than reduced mid-level representations such as chroma.

5.4. METHOD

Our proposed analysis method for spectrogram images is inspired by the eigenfaces method of Turk and Pentland [Turk and Pentland, 1991], which was originally proposed in the context of human face recognition. Since human faces share many common features, by applying Principal Components Analysis (PCA) on a dataset of aligned facial images, a set of basis images ('eigenfaces') can be found, explaining most of the variability found in the face dataset. While PCA has previously been applied as a tool in musical performance analysis [Repp, 1998], this analysis was performed on annotation-intensive IOI data. In contrast, our analysis considers information which only requires alignment of different fragments (as will described in Section 5.5), but no further manual annotation effort.

We apply the same principle to a set of N spectrogram images for a time-aligned music fragment, as represented by N different recordings. Each spectrogram image $\mathbf{x}$ is $(i \cdot j)$ pixels in size. We treat each pixel in the image as a feature; as such, $\mathbf{x}$ is a vector of length $i \cdot j$. We collect all spectrogram images in an $(N \times (i \cdot j))$ matrix $\mathbf{X}$.

By applying PCA, we decompose $\mathbf{X}$ into an $(N \times N)$ matrix of principal component loadings $\mathbf{W}$ and an $((i \cdot j) \times N)$ matrix of principal components scores $\mathbf{T}$. $\mathbf{X}$ can be reconstructed by performing $\mathbf{X} = \mathbf{T} \cdot \mathbf{W}^T$.

Since the PCA is constructed such that principal components are ordered in descending order of variance, dimension reduction can be applied by not using the full $\mathbf{T}$ and $\mathbf{W}$, but only the first L columns of both.

The component scores in $\mathbf{T}$ can now be interpreted and visualized as basis images, each representing a linear component explaining part of the variability in the dataset.

5.5. EXPERIMENTAL SETUP

Unfortunately, no standardized corpora on multiple performances of the same orchestra piece exist.[1] Furthermore, no clear-cut ground truth exists of performance similarity. We therefore consider a dataset collected for the PHENICX[2] project, consisting of 24 full-length recordings of Beethoven's Eroica symphony, as well as 7 recordings of the Alpensinfonie by Richard Strauss. In the Beethoven dataset, 18 different conductors and 10 orchestras are featured (with a major role for the recording catalogue of the Royal Concertgebouw Orchestra (RCO)), meaning that the same conductor may conduct multiple orchestras, or even the same orchestra at different recording moments. While metadata and audio content are not fully identical, in two cases in the dataset (Harnoncourt, Chamber Orchestra of Europe (COE) 1990 and 1991; Haitink, London Symphony Orchestra (LSO) 2005 (× 2)), there are suspicions that these near-duplicates pairs consider the same original recording. In the Strauss dataset, 6 conductors and 6 orchestras are featured: Haitink conducts both the RCO and LSO, and the RCO is represented once more with Mariss Jansons as conductor. The oldest (Mengelberg, RCO, 1940) and newest (Fischer, RCO, 2013) recordings are both featured in the Beethoven dataset.

We will demonstrate insights from the PCA spectrogram analysis in two ways: (1) by

[1] While a dataset of orchestral recordings with multiple renditions of the same piece was used in [Bello, 2011], these recordings are not publicly available.

[2] `http://phenicx.upf.edu`, accessed November 4, 2015.

highlighting several analysis examples in detail in Section 5.6, based on manual selection of musically relevant fragments and (2) by discussing generalization opportunities in Section 5.7, based on aggregation of 4-bar analysis frames.

In both cases, a similar strategy is taken: first, a musical fragment is designated, for which all recordings of the piece should be aligned. Alignment is performed automatically using the method described in [Grachten et al., 2013]. Then, the audio fragments, which are all sampled at Fs = 44.1 kHz, are analyzed using a Hann window of 1024 samples and a hop size of 512, and the corresponding magnitude spectrum is computed using the Essentia framework [Bogdanov et al., 2013]. Combining the spectra for all frames results in a spectrogram image. To ensure that all images have equal dimensions, a constant heigth of 500 pixels is imposed, and the longest fragment in terms of time determines a fixed width of the image, to which all other spectrograms are scaled accordingly. While all recordings are offered at 44.1 kHz, the original recordings sometimes were performed at a lower sampling rate (particularly in more historical recordings). Therefore, a sharp energy cut-off may exist in the higher frequency zones, and for analysis, we try to avoid this as much as possible by only considering the lower 90% of the image. In general, by using raw spectrogram images, a risk is that recording quality is reflected in this spectrum; nonetheless, in the next sections we will discuss how musically relevant information can still be inferred.

5.6. CASE STUDY

In this case study, to illustrate the information revealed by PCA analysis, we will look in detail at information obtained on two selected fragments: the start of the first movement of the Eroica symphony, first theme (bars 3-15), and the 'maggiore' part of the Eroica symphony, second movement (bars 69-104).

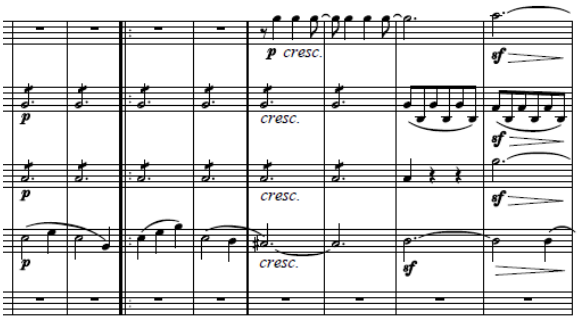


Figure 5.2: Eroica 1st movement, score bars 3-10

5.6.1. EROICA FIRST MOVEMENT, BARS 3-15

A score fragment for bars 3-10 of the first movement of the Eroica is given in Figure 5.2. In our case, we consider the full phrase up to bar 15 in our analysis.

The first three basis images (component scores) resulting from PCA analysis are shown in Figure 5.3. The first component of the PCA analysis gives a smoothed 'ba-

sic' performance version of the fragment. For this very general component, it is rather hard to truly contrast performances. However, a more interesting mapping can be done in higher-order components. As an example, Figure 5.4 displays a scatter plot of the second and third principal component loadings for this fragment.

While as expected, several historical (and acoustically noisy) recordings cause outliers, by comparing the component scores and loadings to corresponding data samples, we still note interpretable differences. For example, the RCO recordings of Fischer and Haitink, of which respective spectrogram images for the excerpt are shown in Figure 5.5, have contrasting loadings on the third PCA component. Judging from the principal component image in Figure 5.3, this component indicates variability at the start of the fragment (when the celli play), and in between the fragments highlighted by the second component; more specifically, a variability hotspot occurs at the sforzato in bar 10. When contrasting two opposite examplars in terms of scores, such as Fischer and Haitink, it can be heard that in the opening, Haitink emphasizes the lower strings more strongly than Fischer, while at the sforzato, Haitink strongly emphasizes the high strings, and lets the sound develop over the a-flat played by violin 1 in bar 10. Fischer maintains a 'tighter' sound over this sforzato.

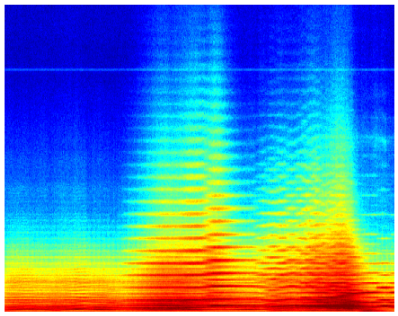
5.6.2. EROICA SECOND MOVEMENT, MAGGIORE

To illustrate findings on another manually selected, slightly longer fragment, we now consider the 'maggiore' part of the second movement of the Eroica. Analyses of scatter plots and component images show that the second principal component is affected by historical recording artefacts. However, this is less so for the third and fourth component, of which the scatter plot is displayed in Figure 5.6. It can be seen that the suspected near-duplicates of Harnoncourt's two COE recordings have near-identical loadings on these components. Next to this, another strong similarity is noted between the recordings of Jochum with the RCO in 1969 and 1978. While these both recordings acoustically are clearly different and also seem to be explicitly different interpretations, there still are consistencies in Jochum's work with the same orchestra for these two recordings.

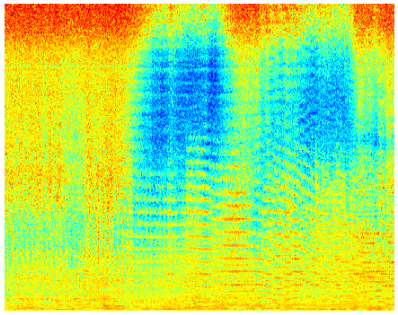
5.7. CORPUS-WIDE CLUSTERING

As demonstrated in the previous section, PCA analysis can be used as an exploratory tool to reveal differences between selected fragments in recordings. However, selecting incidental manual examples will not yet allow for scalable analysis of information over the full timeline of a piece. To do this, instead of pre-selecting designated fragments, we perform a 4-bar sliding window PCA analysis on full synchronized recordings, where bar boundaries are obtained through the score-to-performance mapping obtained in the alignment procedure. Instead of examining individual component images, in each 4-bar analysis frame, we consider vectors of component loadings for the minimum amount of components required to explain 95% of the variance observed. From these component loading vectors, we compute the Euclidean distance between recordings within a frame, and aggregate these at the recording track level.[3]

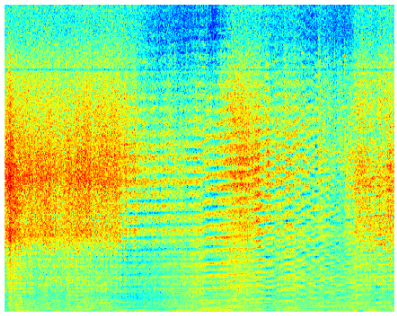
[3] Note that component loadings obtained for different frames cannot directly be averaged, as the components differ per frame. However, observed distances between recordings still remain valid and can be aggregated.



(a) First component



(b) Second component



(c) Third component

Figure 5.3: Beethoven's Eroica symphony, 1st movement, 1st theme start (bars 3-15); first three principal component images.

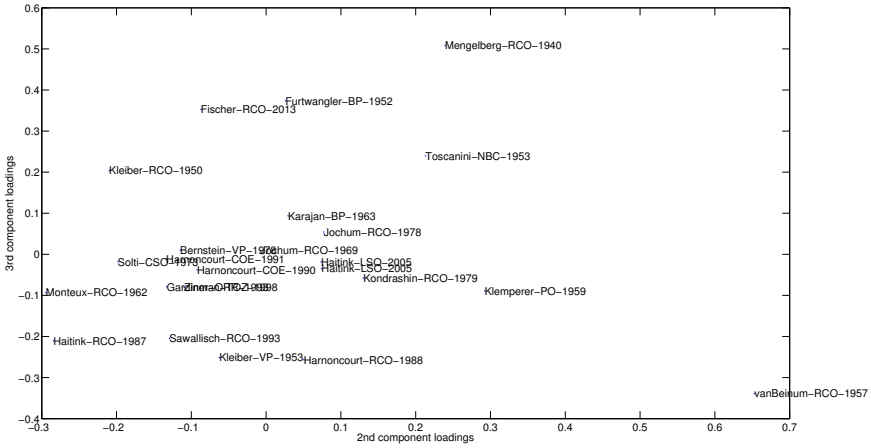


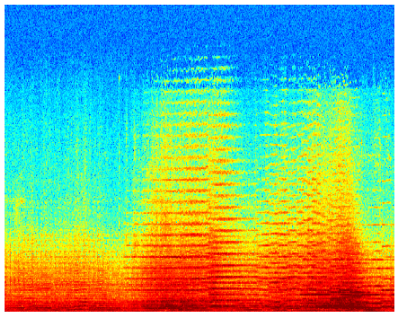
Figure 5.4: 2nd and 3rd PCA component scatter plot for Eroica 1st movement, 1st theme start, bars 3-15.

5

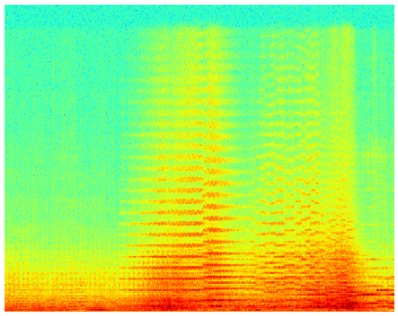
Based on distances found between performances, clustering can be performed. This reveals whether stable performer clusters can found for different movements within a piece, and to what extent clusterings found in local fragments match those found for a full piece.

Regarding the first question, for each of the Eroica movements, we calculated the average between-performer distances per movement, and then made 5 clusters of performers based on Ward's linkage method [Ward Jr, 1963]. While space does not allow a full cluster result report, several clusters co-occur consistently:

- The two Harnoncourt COE recordings consistently form a separate cluster. These are highly likely to be duplicate recordings.
- Haitink's two LSO recordings also consistently co-occur, and like Harnoncourt are highly likely to be duplicate recordings. However, Bernstein's 1978 Vienna Philharmonic recording co-occurs with these two Haitink recordings in the first three Eroica movements, and thus may be similar in terms of interpretation. It is striking that Haitink's 1987 recording with the RCO never co-occurs in this cluster.
- In the first three movements, a consistent cluster occurs with recordings by Klemperer (Philharmonia Orchestra, 1959), Toscanini (NBC Symphony Orchestra, 1953) and Van Beinum (RCO, 1957). While this may be due to recording artefacts, other historical recordings (e.g. Kleiber, RCO 1950 / Vienna Philharmonic 1953) do not co-occur.
- Surprisingly, Gardiner's historically informed recording with the Orchestre Révolutionaire et Romantique (1993) clusters with Kleiber's 1950 RCO recording for the first and last movement of the Eroica. Upon closer listening, Gardiner's choice of concert pitch matches the pitch of Kleiber's recording, and the sound qualities of



(a) Fischer, RCO, 2013



(b) Harnoncourt, RCO, 1988

Figure 5.5: Spectrogram image examples for Fischer and Haitink interpretations of Eroica 1st movement, bars 3-15.

the orchestras are indeed similar (although in case of Kleiber, this is caused by recording artefacts).

- The 1969 and 1978 Jochum recordings with the RCO always co-occur, though in the largest cluster of recordings. As such, they are similar, but no clear outlier pair compared to the rest of the corpus.

Regarding consistent clusterings over the course of a piece, we further illustrate an interesting finding from the Alpensinfonie, in which we compare a clustering obtained on 18 bars from the 'Sonnenaufgang' movement to the clustering obtained for average distances over the full piece, as visualized in the form of dendrograms in Figure 5.7. As can be noted, the clusterings are very close, with the only difference that within the 'Sonnenaufgang' movement, Karajan's interpretation is unusually close to Järvi's interpretation, while Haitink's interpretation is unusually different.

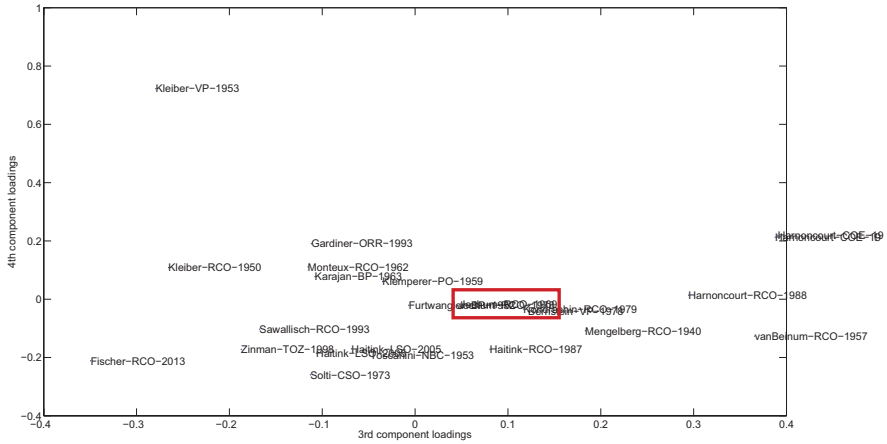


Figure 5.6: 3rd and 4th PCA component scatter plot for Eroica 2nd movement, maggiore. Jochum's 1969 and 1978 recordings occur within the marked rectangular border.

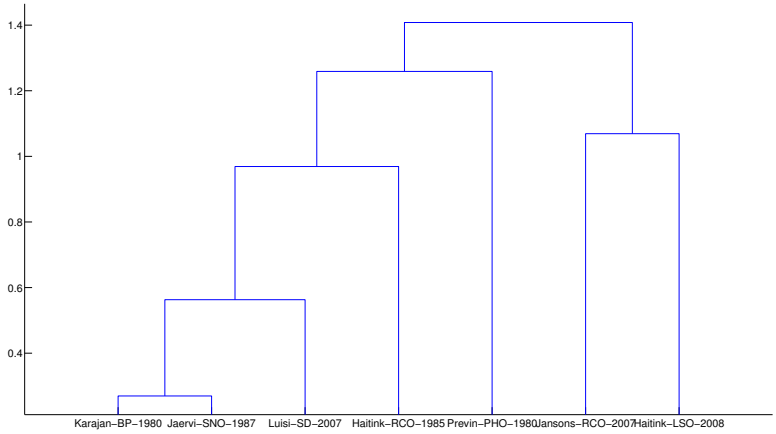
5.8. CONCLUSION

In this chapter, we proposed to analyze differences between orchestral performance recordings through PCA analysis of spectrogram images. As we showed, PCA analysis is capable of visualizing areas of spectral variation between recordings. It can be applied in a sliding window setup to assess differences between performers over the timeline of a piece, and findings can be aggregated over interpretations of multiple movements. While spectrograms inevitably have sensitivity to recording artefacts, we showed that near-duplicate recordings in the corpus could be identified, and historical recordings in the corpus do not consistently form outliers in the different analyses.

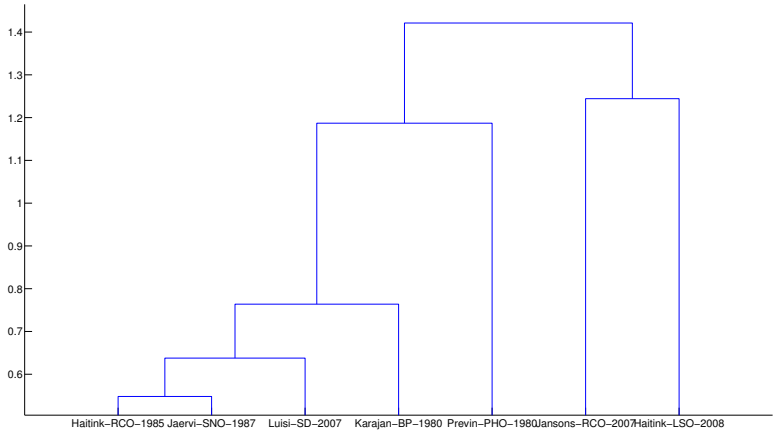
While certain interesting co-occurrences were found between recordings, no conclusive evidence was found regarding consistent clustering of the same conductor with different orchestras, or the same orchestra with different conductors. This can either be due to interference from artefacts and different recording setups, but at the same time may suggest that different conductors work differently with different orchestras.

Several directions of future work can be identified. First of all, further refinement regarding the generation and analysis of the spectrogram images should be performed. At the moment, given the linear way of plotting and high sample rate, the plain spectrogram may be biased towards higher-frequency components, and risks to be influenced by sharp frequency cut-offs from lower original recording sample rates.

Furthermore, it would be interesting to study more deeply if visual inspection of spectrograms can indeed assist people in becoming more actively aware of performance differences. While the spectrogram images are expected to already be understandable to potential end-users, appropriate techniques should still be found for visualizing differences between multiple performers in a corpus. In the current chapter, this was done



(a) 'Sonnenaufgang' fragment (bars 46-63).



(b) Average over full Alpensinfonie.

Figure 5.7: Dendrogram images for performer distances in the Alpensinfonie.

with scatter plots and dendrograms, but for non-technical end-users, more intuitive and less mathematically-looking visualizations may be more appropriate.

One concern that may come up with respect to our work, is that it may be hard to fully associate our reported findings to expressive performance. As indicated, recording artefacts are superimposed on the signal, and effects of different halls and choices of orchestra instruments and concert pitch may further influence acoustic characteristics, which will in turn influence our analysis. Furthermore, since we are dealing with

commercial recordings, we are dealing with produced end results which may have been formed out of multiple takes, and as such do not reflect 'spontaneous' performance.

However, our main interest is not in analyzing performance expressivity per se, but in providing novel ways for archive and search engine exploration, and making general sense of larger volumes of unannotated performance recordings. In such settings, the data under study will mostly be produced recordings with the above characteristics. For this, we believe our approach is useful and appropriate, offering interesting application opportunities.

III

SOUNDTRACK SUGGESTION FOR USER-GENERATED VIDEO

OVERVIEW

Nowadays, not only an increasing amount of digital music is becoming available, but also an (even more rapidly) increasing amount of user-generated video content. This content is not professionally produced, typically recorded with lower-end recording devices such as smartphones in ad hoc settings without a pre-set script. As a consequence, while interesting to the original capturer, this video content usually is not attractive or interesting to broader audiences.

As a consequence of this notion, at the 20th ACM Multimedia Conference (2012), an industrial 'Grand Challenge' was proposed by Google calling participants to contribute ideas "to auto-suggest a cool soundtrack to a user-generated video". More specifically, the challenge was posed as follows:

> *"You have shot a few family videos on your smartphone, but you don't want to upload them to YouTube because they look boring. What if you could find a matching soundtrack? Wouldn't it improve the appeal of the video and make you want to upload it (think of Instagram app [sic] for pictures)?*
> ***Goal:** make a video much more attractive for sharing by adding a matching soundtrack to it."*

When contemplating this problem, we realized that a direct audio-to-video signal-based solution would not be practical. First of all, this was because of expected discrepancies between user-generated content and professionally produced content in the signal domain (both in terms of signal quality and stylistic production features). Furthermore, the intuition was that *imposed narrative* by a user would be more important to the intended message of the video than the actual video content. To deal with this, we proposed a solution which asks a user for a non-musical narrative description of the intended message of the video, and suggests production music based on this description and connected cinematic conventions.

Our investigations into connections between non-musical narrative and music content grew much larger than originally envisioned, heavily integrate existing thoughts from Musicology and Media studies, and still are continuing beyond the scope of this thesis. In this part, in two chapters, which are based on earlier publications, we will reflect on advances made so far in these directions:

- In Chapter 6, we propose the prototypical MuseSync system embodying our original solution to the posed Grand Challenge, and discuss how information from collaborative web resources can be used to exploit common connotative associations between non-musical narrative elements and music descriptions.

- In Chapter 7, through a crowdsourcing study, we investigate in more detail what kinds of non-musical narrative descriptions are invoked by production music, to

what extent these relate to existing theories on the usage of film music, and what elements in the narrative descriptions are stable across subjects.

6

MUSESYNC: STANDING ON THE SHOULDERS OF HOLLYWOOD

Adding music soundtracks to user-generated videos can make such videos more attractive for sharing on the Web. In this chapter, we analyze this use case in more detail and propose a novel story-driven approach to it. In our approach, users first provide their perspective on a video in the form of a short free-text description, together with several supporting keywords. Employing traditional text retrieval methods, we use this input to look for cross-modal multimedia associations as established in cinematic mass media. This information is mined from collaborative online metadata resources on movies and music and enables a first selection of thematically suitable songs from a royalty-free production music database. The selection is reranked based on audiovisual signal synchronization criteria, for which the feature choices follow outcomes of a user survey on cross-modal associations between music and video narrative. The approach is demonstrated in the MuseSync prototype system.

Our approach for MuseSync departs from the traditional approach to study digital music from a positivist viewpoint, focusing on general 'objective' music descriptors. Instead, we strive to put music in a more social and cultural context, leading to novel ways to unify data analysis methods with thoughts from the humanities on musical meaning and significance. This opens opportunities for further work in this area, cross-disciplinary collaborations, and novel contextually oriented music information retrieval application scenarios.

This chapter is an updated, extended and merged version of two previously published works: [1] Cynthia C. S. Liem, Alessio Bazzica, and Alan Hanjalic. MuseSync: Standing on the Shoulders of Hollywood. In *Proceedings of the 20th ACM International Conference on Multimedia — Multimedia Grand Challenge,* pages 1383–1384, Nara, Japan, October 2012 and [2] Cynthia C. S. Liem. Mass Media Musical Meaning: Opportunities from the Collaborative Web. In *11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 689–696, Plymouth, UK, June 2015.

6.1. INTRODUCTION

Over the last years, an explosive growth has occurred regarding web video. With the popularity of smartphones, it is easier than ever to capture informal moments on the go and share these online. However, videos of such informal moments are often not particularly appealing to outsider audience. This situation can be improved by adding a suitable accompanying music soundtrack. The YouTube web video service currently provides a possibility for this in its video manager, which allows a user to search for audio tracks and add these to uploaded videos. However, the music track search can only be done based on song titles. This means that a user should already have a concrete idea of the music he wants to look for, as well as knowledge of the vocabulary used in corresponding music titles. This situation is neither realistic nor user-friendly.

In this chapter, we propose a novel approach to this problem, based on the assumption that a person capturing user-generated video has a clear mental image of what he wants to record, but that the resulting video content on its own only weakly represents this mental image. In order to strengthen the representation and make it truly enjoyable multimedia, suitable soundtrack music should be found for the video. The user should not have to describe this in the form of specialized musical terminology or dedicated song title vocabulary. Instead, we employ a free-text story-driven approach, asking users for a textual plot description and several supporting keywords for their video, after which we automatically compare the provided user input to cinematographic plot situations. Folksonomic song descriptors that are commonly associated with existing plot situations are assumed to also be suitable for the given input video, if the intended story of the video resembles that of the cinematic plot. This results in a pre-selection of candidate soundtrack songs, which is reranked by analyzing audiovisual synchronization matches. In this, we base our audiovisual signal feature choices on the outcomes of a recent user survey[1] which investigated associations between music and video narrative.

This approach has two major novel search engine-related advantages at the user side:

- A user expresses a musical information need to the search engine implicitly through a connected natural language story, which is more natural and straightforward than having to come up with dedicated technical musical terms corresponding to intended music items.

- It is possible to impose different narratives on the same video scene. This gives room to interpret the same visual content in different user-specified ways, which allows for more flexibility and creativity at the user side[2].

The chapter is structured as follows. First of all, we give an overview of existing similar work and related literature. Subsequently, we describe a few practical issues that come with the given use case, and show how this influenced our chosen approach. We then proceed by describing the approach, leading to the presentation of the MuseSync demonstration system. Next to this, we will particularly emphasize the way in which

[1] This is the 'initial survey' discussed in Chapter 7, Section 7.3.5 of this thesis.

[2] Considering the prevalence of cat videos, from a visual concept detection viewpoint, the whole genre consists of near-duplicates. However, the story behind a user's cat video is not the same as someone else's cat video, and a user may want to play around with this story, which our system allows.

connotative connections between non-musical narrative terms and musical descriptions can be found by analyzing collaborative web resources, revealing initial findings on found connotative connections. After this, a short conclusion to the chapter is given.

6.2. RELATED WORK

In the automated (multimedia) retrieval domain, several existing works [Feng et al., 2010, Kuo et al., 2005, Stupar and Michel, 2011] also considered existing cinematic productions as examples for associating music to video. However, in all these approaches, audio and video signal features were directly associated with each other, thus implicitly assuming that low-level signal characteristics hold all necessary information for making cross-modal associations. As we will discuss, this is not a realistic assumption in the case of ad hoc user-generated video.

In musicology and psychology, many studies exist on the way music has been used in the context of movies and other forms of multimedia (e.g. [Tagg and Clarida, 2003] [Cook, 1998] [Cohen, 2005]). It has been conjectured, and even shown in small-scale studies, that music can influence the perception of a moving image, and the other way around. A particularly interesting categorisation of functions of music in relation to moving pictures, though not backed up from a data perspective, was made by Lissa [Lissa, 1965], listing that music can be used to indicate movement, stylize real sounds, represent space, represent time, communicate meaning through deformation of sound, provide a commentary external to the narrative, serve as music within the narrative, indicate psychology of actors, provide a source of empathy, function as a symbol, anticipate action, and serve as a formal unifying factor. This was later recategorized by Tagg and Clarida [Tagg and Clarida, 2003], who proposed a sign typology for music involving sonic, kinetic and tactile anaphones (sounds that were metaphorical for sonic, kinetic and tactile sensory qualities), genre synecdoche, in which a small glimpse of an exotic genre will evoke the full exotic context, episodic markers announcing musical episode transitions, and style indicators. Our approach will build forth on these insights, extracting musical meaning and connotative associations through previous mass media usages, and making use of notions on synchronization and meaning transfer when applying audiovisual synchronization.

6.3. PRACTICAL CONSIDERATIONS

When considering the use case of user-generated video, which is possibly recorded using low-end devices, and has the ultimate purpose of being uploaded and shared online, a few practical considerations come up that in our view should be explicitly taken into account.

6.3.1. USER-GENERATED VIDEO

The characteristics of user-generated video are different from those of professionally produced video. User-generated video is typically not recorded with high-end devices, thus affecting the video quality. The resulting material is often uploaded in its raw form, usually only consisting of a single unedited shot, which is much shorter in duration than an edited video would be. Finally, the content in the video has a strong ad hoc nature,

depicting events and environments significant to the person who recorded them, but lacking a pre-planned narrative component which can be emphasized through scripted stylistic capturing techniques (e.g. close-ups). Because of such characteristics, user-generated video is easily considered as boring, plain, and uninteresting by outsiders who did not share in the same event. The characteristics also negatively influence the possibilities for 'traditional' automated visual signal analysis techniques, such as shot-by-shot analysis techniques or aesthetic and stylistic visual analysis.

6.3.2. LEGAL MATTERS

With this use case, another major issue comes up: copyright law. While this issue is not frequently addressed in multimedia academic publications, it is critical regarding the advancement possibilities for the field. With cinematic productions being under copyright, audiovisual signal data from such productions can only be obtained, accessed and shared within the limits of individual ownership and licenses. Therefore, without licensing deals with large industrial stakeholders (which academic labs usually do not have), it is extremely hard to obtain sufficiently representative datasets in terms of content, scale and reproducibility.

While in our use case, the user will have authoring rights on the video material, explicit care should be taken regarding the choice of soundtrack. In general, resharing an uncleared song in the form of a web video soundtrack can only be done with permission of the rightful owners. Next to this, it is useful to note that even the legal code in all six Creative Commons licenses states that *"[...] where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching" [sic]) will be considered an Adaptation for the purpose of this License."* Thus, even a Creative Commons-licensed song should explicitly allow for adaptations derived from the original, before the song can be legally used for our intended music video creation purposes.

6.3.3. DATA

The considerations above influenced our chosen approach. First of all, we took explicit care to only use legal and public resources for our datasets. Secondly, the user video information is considered to be raw, ambiguous and of non-professional quality, and since it already is hard to obtain sufficiently representative professional visual signal training data, we chose to involve the textual modality with emphasis on imposed underlying narrative as an additional and more concrete source of information.

In our case, we establish cross-modal thematic connections based on textual folksonomic descriptors from two major collaborative web resources: the Internet Movie Database (IMDb) and the last.fm social music service. The IMDb[3] is a large user-contributed resource on movies and TV series, including listings of many types of metadata attributes, including actors and actresses, but also soundtracks and descriptions of plots. last.fm[4] is a social web resource in which users can log and share their music playing behavior, and describe the songs they listen to in tags.

[3] `http://www.imdb.com`, accessed November 4, 2015.
[4] `http://www.last.fm`, accessed November 4, 2015.

To surface cinematic associations between video narrative and music audio, we use IMDb listings of movie soundtracks and movie plots . Through calls to the last.fm API for every listed soundtrack, we crawl social song tags associated with these soundtracks. We retain those movie entries that had both plot information, as well as at least one soundtrack song with last.fm social tags. More details on the types of results obtained through this procedure are given in Section 6.6. For the music, we use royalty-free production music fragments composed by Kevin MacLeod[5], Dan-O[6] and Derek Audette[7], amounting to a dataset of 1084 songs. A large majority of this music is instrumental; in general, in our work we did not consider the textual content of lyrics yet in cases of music with vocals.

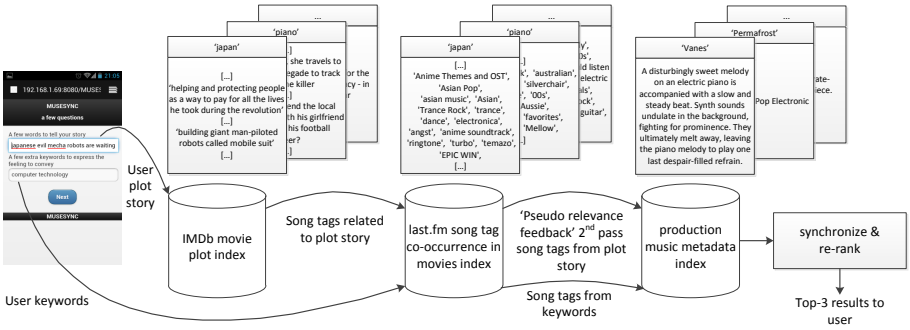


Figure 6.1: Flow scheme of the proposed approach

6.4. PROPOSED APPROACH

In setting up our proposed approach, we first investigated the cross-modal connotation and association power of music in the form of a survey, in which random samples from our production music dataset were played to 49 people from diverse backgrounds. Respondents were asked to describe cinematic and personal scenes to which the music would fit. Important observations from their answers were that people tend to make associations based on temporal developments in the music and the scene (e.g. *'the parts where there are constant harmonies in the background need a more quiet scene than the parts where there is only the beat'*). In addition, when asked about *why* respondents described the scenes they did, a frequent response was that they had heard similar music to similar scenes before.

Our approach builds on these two aspects: association of similar music and similar scenes is done through story-driven soundtrack pre-selection. Then, for these preselected tracks, we use audiovisual synchronization methods to assess cross-modal congruency over the time axis. The synchronization scores will determine the final retrieval ranking. Our approach for the two phases is summarized below, illustrated in Figure 6.1

[5] `http://www.incompetech.com`, accessed July 24, 2012.
[6] `http://www.danosongs.com`, accessed July 24, 2012.
[7] `http://derekaudette.ottawaarts.com/index2.php`, accessed July 24, 2012.

and demonstrated in the MuseSync system, which is made available together with this publication.

6.4.1. STORY-DRIVEN SOUNDTRACK PRE-SELECTION

When starting the MuseSync application, a user is asked to enter a short free-text story description of the video for which he searches a soundtrack, as well as several tag-like keywords describing the intended 'feel' of the video. Subsequently, we employ a traditional text retrieval approach in three steps:

1. We first compare the free-text user description (the 'plot' of the video) to IMDb plot descriptions of existing movies. This is done because we want to associate similar plot elements to similar corresponding music keywords (in our case, last.fm song tags). Therefore, we employ a search index consisting of documents which have music song tags as keys. The contents of a document from this index consist of the concatenated plot descriptions of all movies that have a soundtrack song with the particular music tag forming the document key.

2. Subsequently, we clean up the obtained music song tags through a step similar to pseudo relevance feedback. We employ a song tag co-occurrence search index, which once again has music song tags as keys. A document from the index consists of all music song tags that occur together with the key tag within a movie. Ultimately, we want to end up with n tags that will be used in the final search step. Half of these n tags are obtained by querying the song tag search index with the tags obtained from our first step. The other half is obtained by directly querying the song tag index with the keywords that the user entered. For our current approach, n was chosen to be 50.

3. Finally, we use the song tags resulting from the previous step as a query to a music metadata search index. This index is built based on metadata descriptions of the songs (genre, instrumentation, associated mood, etc.), as entered by the original song composers.

For all search indices mentioned above, we employed the standard search facilities of the Lucene framework, which uses length-normalized tf*idf measures in order to score documents.

6.4.2. VIDEO TO AUDIO SYNCHRONIZATION

The soundtrack pre-selection phase results in a set of candidate tracks that thematically corresponds to the given user description. Both in the psychology (e.g. [Cohen, 2005]) and multimedia (e.g. [Hua et al., 2004]) fields, it has been mentioned that good synchronization or congruence of video and audio segments will enhance the perception of both modalities, and stimulate unification of the meaning perceived for the individual modalities. Therefore, we will base our final soundtrack ranking on audiovisual synchronization.

In the multimedia academic domain, the video component has traditionally been considered as the strongest modality in a multimodal setting. Audio going with this video component is usually considered to be subordinate, resulting in the expectation

that an added audio component shall be modified (typically through time warping) in case of a non-perfect temporal fit to the video. However, as we mentioned before, in the case of user-generated video we do not consider the video component to be a very strong modality. Therefore, in our work we will keep both the video and music soundtrack streams in their original forms and not modify their internal temporal discourse in our synchronization techniques. Instead, for our synchronization we will only shift the signals by a fixed lag, found through cross-correlation.

The synchronization is achieved on audio and video feature vectors that relate to the findings from our survey. Since temporal developments in music (especially at the timbral level) were considered by our respondents to be important, we employ an audio feature capturing this information. For this, we start from the work described in [Foote et al., 2002], where novelty of an audio signal is measured through the analysis of a self-similarity matrix. Audio changes are detected comparing two adjacent regions around a point lying on the diagonal and correlating with a checkerboard kernel. High correlations between the regions and the kernel correspond to adjacent regions with low cross-similarity. Local minima of that series of values well encode “when" a relevant change occurs. We employ this approach on timbral MFCC information, employing a large analysis window of 1 second in order to avoid short-time fluctuation. However, while this approach is effective in detecting audio changes, it does not sufficiently capture the relevance of a change. Therefore, we also run an onset detector on the music signal and take the signal intensity at every detected onset time as a weighting measure. Multiplying these onset intensities with the novelty curve, audio changes that go together with strong accents will get emphasized.

Regarding the video signal, we assume to have a single raw shot, meaning that no sudden cuts or cross-fades are present, although one might observe fast camera motions like panning. Since user-generated videos can have very diverse visual content, we cannot concentrate on a specific class of objects or dedicated concept detectors, but only can consider general descriptors. We therefore employ the MPEG-7 motion activity descriptor presented in [Jeannin and Divakaran, 2001], for which an existing user study reported that the descriptor was capable of capturing ‘intensity of action’. An attractive feature of this descriptor is that it is not requiring heavy computation: for each video frame, the motion activity score is given by the standard deviation of the magnitudes of the motion vectors which can be directly read from the encoded video stream.

By computing synchronization (cross-correlation) scores between the given video and all suggested soundtracks, we obtain a soundtrack ranking. Since we assume that multiple soundtracks can fit well to the video content and the given user description, the user will not just be provided with the single best soundtrack, but rather receives synchronized results for the three best-scoring soundtracks. A more extensive description of this module can be found in [Bazzica, 2012].

6.5. FURTHER SYSTEM DETAILS

Having described the conceptual flow and methodological approaches of the MuseSync system, in this section we will give some implementation details. A system component diagram is given in Figure 6.2. As can be seen from the diagram, the system has a user-facing web front-end. At the client side, jQuery mobile and HTML5 are used, allowing

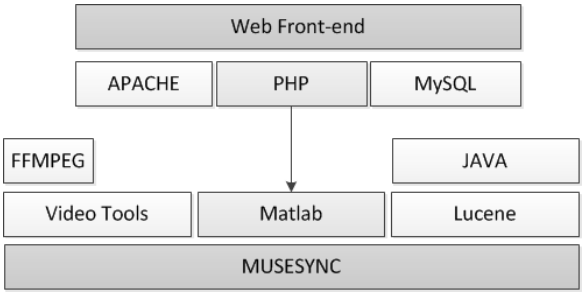


Figure 6.2: MuseSync system components

MuseSync to be accessible as a web app in the browser via any platform that allows for local file browsing (which is necessary for selecting a video). This means that MuseSync is accessible to both desktop and mobile platforms that allow for this functionality, including Android. We explicitely did not choose to make the full system run natively on a mobile device, but rather prefer to do the music soundtrack selection and synchronization server-side. This is done so the user does not need to have a full local music dataset on every device from which he wants to use the system, and to facilitate use scenarios with separate music repositories, which can be individually maintained by the responsible owners.

At the server side, Apache and PHP are used, interfacing with Matlab, which is called as an external process through a PHP script. The visual feature descriptor is extracted with an FFmpeg-based tool written in C, allowing for fast reading of motion vector streams. The audio feature descriptor is extracted through the Matlab MIRtoolbox. As mentioned earlier, Lucene is used for handling the textual user input.

While our approach has mainly been designed for instrumental production music with composer-contributed metadata annotations, this does not mean that a user will not be able to use personal music collections at all in the MuseSync system. As long as it is possible to retrieve meaningful textual descriptors for a song (which can be obtained from services like last.fm), the song can be handled by the system.

6.6. THE POTENTIAL OF COLLABORATIVE WEB RESOURCES: A CLOSER LOOK

In this section, we will take a closer look at the types of connotative associations between musical and non-musical narrative elements that can be made by connecting information collaborative web resources.

As mentioned before, we first crawled the IMDb for movies which also had soundtrack listings associated to them. For each of these soundtracks, we queried last.fm for taggings of soundtrack songs, and kept those movies that had at least one soundtrack with a last.fm tag associated to it. 22,357 unique IMDb movies with plot descriptions had at least one soundtrack song with a last.fm tag associated to it; in total, considering the soundtracks of all these movies, 265,376 song tags could be found. Subsequently, we addressed two questions:

- *Do different narrative elements emerge for different song tags?* For this, a mapping was made from song tags to all movie plots which had a soundtrack to which this song tag was associated. For example, the song tag 'guitar' will be mapped to all plot descriptions in the corpus of movies which had at least one soundtrack song on which 'guitar' was used as a last.fm tag.
- The other way around, *do different song tags emerge for different narrative elements?* This is the case that actively is integrated as part of the MuseSync system. As indicated above, we built an Apache Lucene[8] search index, framing our data in an automated text retrieval scenario. Using the map from song tags to movie plots, we treat the song tag as the document key, and all corresponding movie plot stories as document text. Subsequently, we can issue natural language queries (in MuseSync: user story descriptions) to the search index, which will surface the 'documents' matching the query best (so, cases in which movie plot descriptions cause a close match to the issued query), and then consider the corresponding document keys, which consist of song tags. To reduce noise, in the results reported in this section, we only index song tags which were used at least 1000 times in last.fm.

6.6.1. INITIAL OUTCOMES

In this subsection, we report initial outcomes resulting from exploration of the data connections which were established as explained above. Unfortunately, no standard metrics or ground truth rules exist on the relation between music song tags and narrative elements. Therefore, as a first way to still display emergent patterns in the data, in discussing outcomes we will use Wordle[9] word cloud visualizations, which apply common automatic statistical text analysis methods to visualize the most important words in text corpora.

DO DIFFERENT NARRATIVE ELEMENTS EMERGE FOR DIFFERENT SONG TAGS?

We take the mapping from song tags to movie plots as described in the previous section, and then examine what kinds of words occur in the aggregated collection of movie plots associated to a certain song tag. In Figure 6.3, word clouds are displayed based on several song tag queries reflecting different music genres.

Like any other collaborative web resource, both the last.fm and IMDb corpora are noisy. Furthermore, especially when many movie plots are found for a song tag, while the text analysis provided by Wordle filters out common stopwords and highly frequent words, certain universal elements which are no stopword, but still occur in almost all movie plot narratives ('one', 'life', 'man', 'woman') stand out. More sophisticated language analysis would be needed to filter those out.

Still, some distinctive characteristics already emerge for different genres. For example, while 'family' occurs in each of the word clouds, the movie plots associated to 'opera' have a stronger connection to family-oriented themes like ' marriage' and 'Christmas'. Movie plots associated to 'salsa' have an explicit relation to dancing. Movie plots associated to 'rap' suggest that younger main characters are involved ('school' is more prevalent than in other tags) and seem to suggest slightly stronger male connotations than

[8] `lucene.apache.org`, accessed November 4, 2015.
[9] `www.wordle.net`, accessed November 4, 2015.

the other word clouds ('man' and 'father' are relatively large; 'brother' is equally sized to 'girl', but 'woman', 'girlfriend' and 'mother' are clearly smaller; finally, any given names emerging in the word cloud are male names).

DO DIFFERENT SONG TAGS EMERGE FOR DIFFERENT NARRATIVE ELEMENTS?
In the reversed scenario of looking up collections of song tags for non-musical narrative elements, further interesting socially established aspects can be found. Figure 6.4 shows what song tags are returned by querying the Lucene search index using several names of cities. Again, certain very popular song tags occur for all word clouds (in particular 'rock', the most frequently used tag on last.fm). However, it is clear that beyond this, differences occur between the word clouds, revealing connections between geographical locations and typically associated music styles (e.g. 'chanson' for Paris, 'blues' for Chicago, 'anime' for Tokyo). In Figure 6.5, we show further interesting results for various queries which are non-musical, but do express a narrative context. Again, the broadly popular 'rock' song tag strongly occurs in all results, but next to this, we notice different nuances which indeed represent typical connotations for the given queries (e.g. some 'rougher' genres like metal and punk for 'car chase', genres associated to warmer regions like Jamaica, Italy and Brazil for 'beach holiday', and some less rough genres like instrumental, atmospheric, blues and classical for 'candlelight dinner').

6.7. CONCLUSION

In this chapter, we presented a novel approach to soundtrack retrieval for user-generated videos, exploiting cinematic knowledge through story-driven folksonomic text retrieval, building on knowledge in large collaborative web resources. Due to the nature of user-generated video content, we do not follow the common approach of direct audio-to-video signal matching. Instead, we first ask explicit user input. However, instead of soliciting an explicitly musical query, we target a more natural and more flexible way of implicit querying by providing a desired narrative for the video. Subsequently, the narrative is connotatively connected to musical terms by using connections obtained from the collaborative web resources. Information on narrative movie plot elements and soundtrack use was obtained from the IMDb, while information on descriptive song tags for soundtrack songs was obtained from last.fm. As initial data visualizations reveal, connotative associations between music tags and narrative plot elements indeed are reflected in these collaborative web resources. As a final step, audiovisual synchronization is applied for reranking music matches, based on high-level features described by users.

Illustrative concrete results of the MuseSync system, the current system implementation, as well as other supporting material are available at `http://multimediaeval.org/musesync/`.

Future work involves implementing more advanced text analysis mechanisms for analyzing the input narrative and making the internal connections between resources, as well as larger-scale evaluations of the system with users. As our initial visualizations reveal, culturally specific notions are encoded in collaborative web resources, opening doors to models and algorithms to infer cultural views on musical meaning from data. Given the strong rooting of our approaches in literature from the Humanities and Social Sciences, we believe concrete possibilities for cross-disciplinary cooperations between

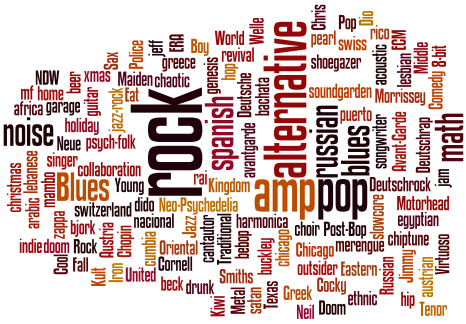
(a) ‘opera’

(b) ‘salsa’

(c) ‘rap’

Figure 6.3: Wordle word clouds of aggregated movie plots for several genre-related music song tags.

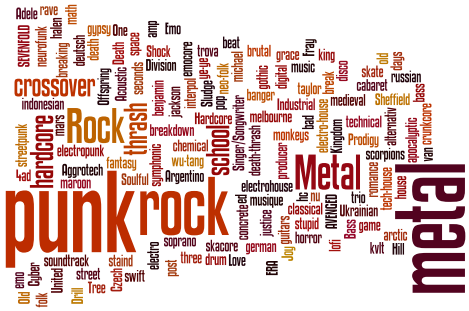
(a) ‘Paris’



(b) ‘Chicago’

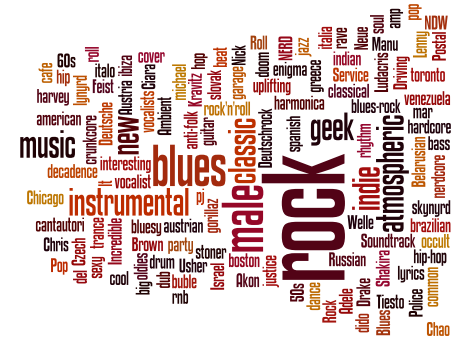
(c) ‘Tokyo’

Figure 6.4: Wordle word clouds for song tags retrieved for city name queries on the Lucene search index.



(a) ‘car chase’

(b) ‘beach holiday’



(c) ‘candlelight dinner’

Figure 6.5: Wordle word clouds for song tags retrieved for various other non-musical narrative context queries on the Lucene search index.

computer (data) science and these fields can suitably be further investigated in the context of this work, as well as novel querying paradigms for Music Information Retrieval systems based on cultural context.

7

WHEN MUSIC MAKES A SCENE — CHARACTERIZING MUSIC IN MULTIMEDIA CONTEXTS VIA USER SCENE DESCRIPTIONS

Music frequently occurs as an important reinforcing and meaning-creating element in multimodal human experiences. This way, cross-modal connotative associations are established, which are actively exploited in professional multimedia productions. A lay user who wants to use music in a similar way may have a result in mind, but lack the right musical vocabulary to express the corresponding information need. However, if the connotative associations between music and visual narrative are strong enough, characterizations of music in terms of a narrative multimedia context can be envisioned, as was proposed in the MuseSync system in the previous chapter. In this chapter, we present the outcomes of a user study considering this problem. Through a survey for which respondents were recruited via crowdsourcing methods, we solicited descriptions of cinematic situations for which fragments of royalty-free production music would be suitable soundtracks. As we will show, these descriptions can reliably be recognized by other respondents as belonging to the music fragments that triggered them. We do not fix any description vocabulary beforehand, but rather give respondents a lot of freedom to express their associations. From these free descriptions, common narrative elements emerge that can be generalized in terms of event structure. The insights gained this way can be used to inform new conceptual foundations for supervised methods, and to provide new perspectives on meaningful and multimedia context-aware querying, retrieval and analysis.

The contents of this chapter previously were published as Cynthia C. S. Liem, Martha A. Larson, and Alan Hanjalic. When Music Makes a Scene — Characterizing Music in Multimedia Contexts via User Scene Descriptions. *International Journal of Multimedia Information Retrieval*, 2(1):15–30, 2013.

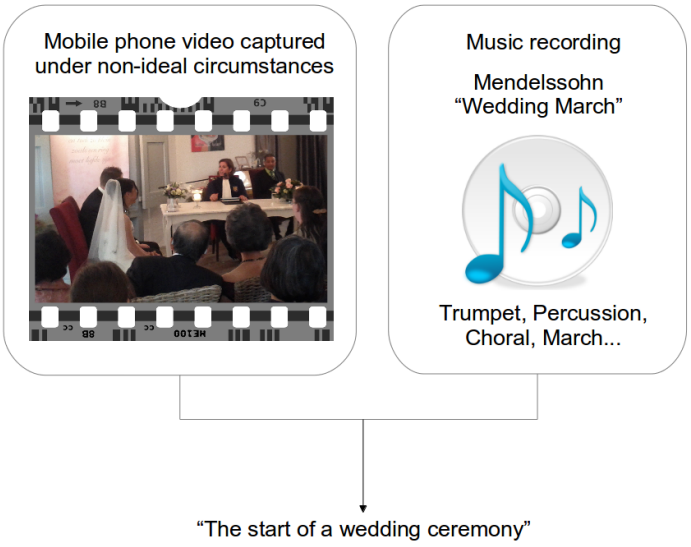


Figure 7.1: Certain social events have grown to imply certain types of music and vice versa. This can be exploited to clarify the message of multimedia productions.

7

7.1. INTRODUCTION

Music is used in different ways. Besides serving active listening purposes, it can function as background entertainment alongside everyday human activities. Next to this, it also frequently occurs as an important reinforcing and meaning-creating element in multimodal human experiences. Many social occasions, ranging from parties to more ceremonial events such as weddings and funerals, would not be the same without music. In many digital audiovisual productions, music plays an essential role as well, being exploited for setting the atmosphere of a visual scene, characterizing key characters in a scene, implying situations that are not directly visible to the audience, and much more.

In the present-day Web era, creating and sharing such productions is no longer the exclusive privilege of professional creators, as can be seen from the continuously rising usage numbers of video sharing sites such as YouTube[1]. Everyday users are increasingly interested in capturing personally significant situations they encounter on the go, and in sharing these with a world-wide audience. Especially if the captured video material would be of low quality, a fitting music soundtrack can greatly impact and clarify the video's significance (see Figure 7.1 for an example).

[1] `http://www.youtube.com/t/press_statistics`, accessed December 13, 2012.

7.1.1. CONNOTATIVE ASSOCIATIONS IN MULTIMEDIA CONTEXTS

In both the cases of enriching real-life social occasions, as well as influencing the message of a digital audiovisual production, music will not be considered in the form of isolated songs. Instead, it is placed in a context that is not necessarily musical itself. The context may be time-dependent, and is composed of *multimodal* information: information which is relevant to different parts of the human sensory system. When music is placed in this context, a situation arises in which a combination of information in different modalities and formats provides added value in comparison to the original context. In other words, music will now be part of a *multimedia context*.

The added value that music provides to a context can work in a *cross-modal* way, in which information relevant to one part of the human system, such as the auditory channel, influences the perception of information relevant to another part of this system, such as the visual channel. In terms of meaning and significance, under these circumstances, music can have influence outside its own domain. As soon as this becomes conventional, a non-musical context may directly get associated with specific types or properties of music, and vice versa. For example, a bugle call typically has grown to evoke militaristic images, while a lullaby rather evokes images of infants. This way, outside of *denotative* meaning dealing with purely musical properties, music also gets a referential *connotative* meaning layer [Meyer, 1968].

Indeed, in multimedia productions, cross-modal connotative effects and associations are heavily exploited. Since the early years of the moving picture, multimedia professionals have been trained to analyze the techniques to achieve this [Kalinak, 1992, Lissa, 1965], and to actively exploit these themselves when creating their own productions [Lang and West, 1920, Prendergast, 1992].

Due to exposure in daily life and mass media, non-professional lay users may be capable of recognizing prototypical associations between music and contextual information in other modalities as well. While it is unrealistic to assume that these users are capable of describing their musical information needs for a desired non-musical context in terms of musical or signal characteristics, it is likely that they can indicate high-level characteristics of the envisioned multimedia result. In fact, since low-level signal properties cannot cover all aspects of meaningful associations, and since the optimal musical match will differ per case, it would even be more realistic to describe an information need in terms of the envisioned multimedia result, rather than in terms of purely musical characteristics. This point is illustrated in Figure 7.2: based on low-level timbral feature properties, it would be hard to distinguish between the three displayed songs (e.g., they all contain trumpets and percussion). However, the associations triggered by the songs are very different. The narrative context of the user query influences the relevance of these associations. For example, 'Duel of the Fates' would be the most appropriate match if the user envisions an epic battle scene. On the other hand, if the user would wish to show the rivalry between his pets in a funny or ironic way, the Wedding March could actually become a more suitable match.

7.1.2. CONTRIBUTIONS AND OUTLINE

In this chapter, we take a novel approach to multimedia context-oriented annotation. Focusing on production use cases, we investigate the feasibility of characterizing music

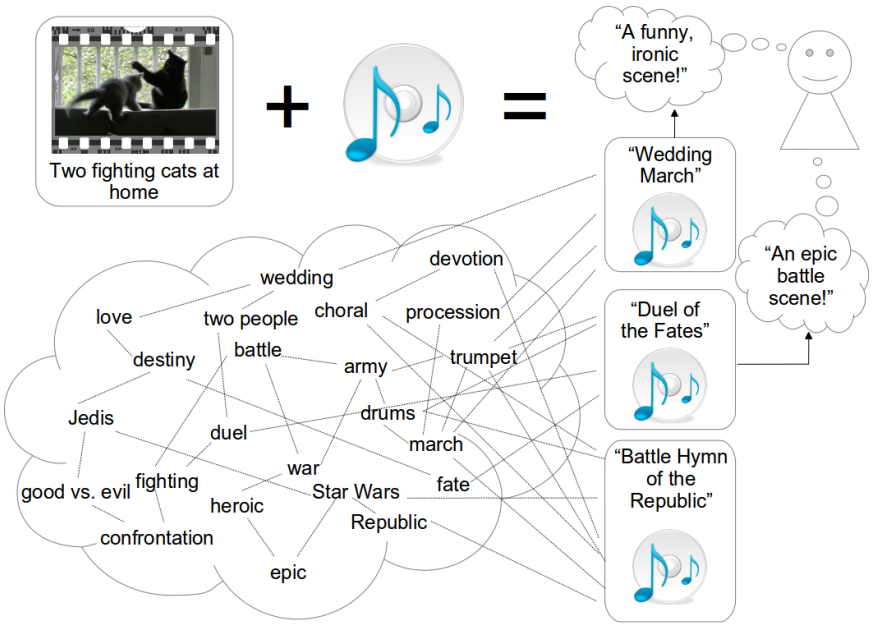


Figure 7.2: An example of potential connotative meaning-formation in the intended multimedia production use case. We have a user video and three potential soundtrack candidates, for which several possible meaning associations are indicated with dashed lines. Due to these associations, the choice of soundtrack will influence a user's interpretation of the resulting multimedia scene.

via high-level descriptions of *its context and intended use* (envisioned multimedia results that fit to it) rather than in terms of *modality-specific content* (musical feature descriptors). We do *not* assume the annotation vocabulary to be known beforehand. Instead, we solicit free-form text responses for our survey, giving full control over the annotation vocabulary and its typical usage to the survey respondents, and inferring the characteristics of the vocabulary and its usage only after the responses are received.

The results of our work push the music and multimedia information retrieval fields forward in two ways:

- As we will show, contextual multimedia scene descriptions can reliably be matched to the music fragments that triggered them. This opens entirely new perspectives on cross-modal connotative annotation and querying strategies.
- The conceptual levels at which descriptions are generalizable, as well as the self-reported reasons for giving the descriptions, provide insights into criteria that caused cross-modal connotative associations to emerge. These can inform the design and realization of supervised automated methods and ground truth at the multimedia signal data level, such that practical systems can be realized that take these associations into account.

This chapter is structured as follows. We start with giving an overview of related

work in Section 7.2. We then present our experimental setup in Section 7.3. We proceed by describing statistics related to the crowdsourcing mechanisms in Section 7.4. Section 7.5 analyses the results of a task testing the recognizability of correspondences between multimedia descriptions and the music fragments that triggered them. Section 7.6 focuses on the analysis of the obtained free-form text survey responses, investigating which narrative elements emerge and at which conceptual levels descriptions are generalizable. Finally, Section 7.7 provides a conclusion and outlook to follow-up opportunities to our findings in the music and multimedia information retrieval fields.

7.2. RELATED WORK

Our work in this chapter is an exploratory user study. In the music information retrieval (Music-IR) domain, the importance of user-centered versus system-centered approaches has attracted attention for several years [Downie et al., 2009, Lee et al., 2012], although actual adoption of user-centered approaches still is rare. Recent overviews on the types of user studies that have been conducted in the field and their impact can be found in Weigl and Guastavino [Weigl and Guastavino, 2011] and Lee and Cunningham [Lee and Cunningham, 2012]. Most of the few existing Music-IR user studies have been focused on information needs, search strategies and query analysis topics (e.g. [Cunningham et al., 2003, Downie and Cunningham, 2002, Lee, 2010]). In our case, while our work will ultimately relate to search and retrieval, we will first and foremost consider a scenario of data description.

In terms of the domain under study, the work of Inskip et al. [Inskip et al., 2008, 2010], dealing with discourse in search and usage of pre-existing commercial music, is close. However, this work has been focusing on professional stakeholders only and was more oriented towards identifying typical language usage in decision-making and the production pipeline.

Looking at previous data description work, well-known human-tagged Music-IR datasets such as CAL500 [Turnbull et al., 2008] and MagnaTagATune [Law and Ahn, 2009] strongly focused at having the tag annotations satisfying inter-annotator agreement criteria, but did not explicitly analyze user motivations behind the resulting annotations. Only very recently, work by Lee et al [Lee et al., 2012] appeared in which the rationale behind assigned musical mood labels was investigated more deeply. In all these cases, music is treated as an isolated information source, not exceeding the earlier mentioned idea of pleasant background diversion accompanying everyday human activities.

The notion of music in *context* has been emergent in recent years. Schedl and Knees [Schedl and Knees, 2009] mention cultural context information, but this context information (such as web pages on performing artists) once again remains largely oriented towards the idea of music in an isolated or accompaniment function. Kaminskas and Ricci [Kaminskas and Ricci, 2012] give an overview of work in different categories of music context, including multimedia. These approaches are strongly data-driven, while deeper insights into the underlying mechanisms establishing context would be desirable.

Recent work considering cross-modal relations between music and contextual, external information includes [Cai et al., 2007, Feng et al., 2010, Kuo et al., 2005, Li and Shan, 2007, Stupar and Michel, 2011]. These either fix the cross-modal association do-

main to be purely affective, or directly match signal features in different modalities to each other. In our work, we will adopt the view from from [Kuo et al., 2005, Stupar and Michel, 2011] on cinema as prototypical multimedia form. However, we want to look beyond pure signals and fixed domains, at *semiotics* rather than *semantics*, and *connotation* rather than *denotation*. Ideas on semiotic hypermedia and multimedia have been pioneered in [Colombo et al., 2001, Nack and Hardman, 2001]. A recent rare work on connotation was done by Benini et al. [Benini et al., 2011], in which annotations in a connotative space was shown to produce higher user agreement than emotional label tagging.

Semiotic aspects have been a topic of interest in the field of musicology for several decades, one of the first pioneers being Nattiez [Nattiez, 1973]. While Nattiez argues that music lacks the communicative power to picture concrete events, it still is capable of suggesting general outlines of these events. In an experiment very similar to the way we will set up our description survey, Nattiez played a piece of classical symphonic program music to secondary school students, asking them to describe the story the piece was telling. While the original program of the piece did not appear, certain categories of stories clearly occurred more frequently than others. In addition, the study showed that not all elements of formal musical structure were being connected to stories, and that less relevant events in relation to the musical argumentation (e.g. a final chord) are nonetheless being picked up as relevant to the story descriptions.

An interesting aspect of present-day musicological thought is that it has shifted away from absolutist, positivist views (assuming an objective truth 'in the music data') towards subjective, contextual and culturally-specific aspects of music [Cook, 1998], up to the point that 'the music itself' is nowadays even considered a taboo concept [Wiering and Volk, 2011]. Because of this view, musicology has strongly been studying contextual topics in which music plays a more significant role than mere accompaniment. Film music and mass media studies also has been a long-time interest in the field. A very useful categorization of film music functions has been made by Lissa [Lissa, 1965], mentioning that music in film can be used to indicate movement, stylize real sounds, represent space, represent time, communicate meaning through deformation of sound, provide a commentary external to the narrative, serve as music within the narrative, indicate psychology of actors, provide a source of empathy, function as a symbol, anticipate action, and serve as a formal unifying factor.

Influenced by Lissa, Tagg [Tagg and Clarida, 2003] proposed a sign typology for music involving sonic, kinetic and tactile *anaphones* (sounds that were metaphorical for sonic, kinetic and tactile sensory qualities), *genre synecdoche*, in which a small glimpse of an exotic genre will evoke the full exotic context, *episodic markers* announcing musical episode transitions, and *style indicators*. In [Tagg and Clarida, 2003], Tagg and Clarida also report findings from a decade-long study, in which association responses to ten different well-known theme and title songs were collected and categorized. Major differences with our approach are that immediate, very short associations were sought (respondents only had a few seconds to write these down) while we solicit more elaborate user input.

Following our vision of free-form text descriptions of multimedia productions, the results will be text descriptions of visual narratives, needing to be analyzed in terms of

narrative structure [Bal, 2009, Barthes, 1977, Chatman, 1980]. In our analyses, we will adopt a simple, pre-theoretical and widely-accepted dictionary definition of a narrative as *"a spoken or written account of connected events; a story"* [OUP, 2008]. This definition motivates us to take events as basic building blocks of a narrative, and turn to event theory for a basic typology of events as they are most naturally conceptualized in human semantic systems. More specifically, we will build upon the work of Vendler [Vendler, 1967] regarding the structure of events in terms of classes of verbs. Four classes are differentiated on the basis of temporal structure: *states* have no internal structure, *activities* involve internal change but no endpoint, *achievements* involve an instantaneous culmination or endpoint and *accomplishments* involve a build up period and then a culmination. Collectively these classes are referred to as event classes. A basic distinction can be made between culminating events (achievements and accomplishments) and non-culminating events (states and processes).

Our chosen methodology for soliciting user responses is that of *crowdsourcing*, an approach which is currently emerging in the Music-IR field, e.g. in [Lee et al., 2012, Mandel et al., 2010, Urbano et al., 2010]. In multimedia information retrieval, previous crowdsourcing work typically focuses on collecting annotations describing the explicitly depicted content of multimedia. For example, Nowak and Rüger [Nowak and Rüger, 2010] employ crowdsourcing to collect information about images and focus on labels for visual concepts. Such a task targets collecting 'objective' annotations, which can be considered correct or incorrect independently of the personal view of the annotator. In contrast, in this work we use crowdsourcing to carry out a highly open-ended and creative task for which no independent gold standard can be considered to exist. In this way, we take a step beyond the work of Soleymani and Larson [Soleymani and Larson, 2010], in which larger tasks than microtasks were already experimented with and reported on, and Vliegendhart et al. [Vliegendhart et al., 2011], in which crowdsourcing tasks requiring a high imaginative load were supported. Because we are interested in eliciting associations made by humans, our use of crowdsourcing also shares commonalities with its use for behavioral studies, as discussed by Mason and Suri [Mason and Suri, 2012].

7.3. EXPERIMENTAL SETUP

7.3.1. INFRASTRUCTURE: AMAZON MECHANICAL TURK

The user studies presented in this chapter were conducted using the infrastructure of the Amazon Mechanical Turk[2] (AMT) crowdsourcing platform. AMT is an online work marketplace on which *requesters* can post *Human Intelligence Tasks* (HITs), which can be taken up by *workers*. Each HIT typically is a microtask, soliciting a small amount of human effort in order to be solved, and yielding a small amount of payment to a worker who succesfully completes it. Our choice for this platform is e.g. supported by findings in previous work by Paolacci et al [Paolacci et al., 2010], in which experiments in the area of human judgment and decision-making were performed both on AMT and using traditional pools of subjects, and no difference in the magnitude of the effects was observed.

A requester designs a HIT template and separately specifies the input data to be used with this. Furthermore, for one HIT, a requester can release multiple *assignments*, re-

[2] `http://www.mturk.com`, accessed December 13, 2012.

Part 1: General questions on the music fragment

1. How familiar are you with this music fragment?
   [*very unfamiliar - somewhat unfamiliar - somewhat familiar - very familiar*]
2. Please characterize the music fragment.
   (a) How positive or negative do you consider the message of this fragment to be?
   [*very negative - somewhat negative - neutral - somewhat positive - very positive*]
   (b) How active or passive do you consider this fragment to be?
   [*very passive - somewhat passive - neutral - somewhat active - very active*]

Part 2: Describe a cinematic scene to which this music fragment can be a soundtrack
Imagine you would be a film director, preparing your greatest piece of cinema work so far. You have a huge budget for this, meaning you could hire a world-class crew to support you, and you don't have any limitations regarding filming locations, costumes, props etc. either. The award-winning soundtrack composer perfectly understands what you want to express with your movie—and came up with this fragment as a musical sample to go with a certain scene in your movie. In the following questions we would like to learn from you what the characteristics of this scene would have been. Please use your imaginative power and be as creative and illustrative as you can.

3. What is happening in the scene?
   [*free-form text input*]
4. If you had any specific film genre in mind of which this scene would be a part, please mention it here.
   [*free-form text input*]
5. Who are the key characters in the scene?
   (a) If you indicated any key characters in your scene description above, please introduce them here (e.g., 'an old woman', 'James, a 40-year old father', 'an alien from Mars', 'a pit bull terrier', 'Napoleon Bonaparte'...).
   [*free-form text input*]
   (b) What do these characters look like? What kind of persons/creatures are they? Are they thinking or feeling anything specific? If there are multiple characters, do they relate to and interact with each other? If so, in what way? (e.g., 'James spends most of his time at the office. He doesn't like that and wishes to spend more time with his kids.', 'The pit bull terrier is evil and very aggressive. He does not get along with the docile cat next door.' etc...)
   [*free-form text input*]
6. What is the general setting of the scene?
   (a) Where does the scene take place? (e.g., 'China', 'in the mountains', 'in a Tuscan Medieval convent')
   [*free-form text input*]
   (b) When does the scene take place? (e.g., 'in 1815', 'in Summer', 'during the Second World War', 'at midnight')
   [*free-form text input*]
   (c) Please give some indications of what the setting visually looks like (e.g., 'The sun is shining brightly over the convent. It is a recently restored building, that attracts a lot of tourists during the Summer season.')
   [*free-form text input*]
7. Please describe why you chose the scene, characters and setting described in the previous questions for the music fragment.
   [*free-form text input*]

Figure 7.3: Questions used for the cinematic scene description survey.

quiring that multiple different workers carry out the same task on the same input data. HITs are typically released in *batches*, grouping the assignments of multiple HITs with the same template. This way, upon completing a HIT, a worker can easily proceed to taking up another HIT of the same type, but with other input data. All completed work has to be reviewed by the requester and approved.

If a worker did not complete a task in a satisfactory way, the requester may reject the work and repost the task. A requester can limit the access to HITs by imposing requirements on the worker's status. Besides requirements based on general worker properties, such as a worker's approval rate on earlier tasks of the requester, a requester can also manage the access to HITs by soliciting and handing out custom qualifications to workers.

In the following two subsections, we will present the core of a description survey and a subsequent music ranking task on which the work in this chapter are based. After this, we proceed by describing the data that was used for our experiments, and specifying our task runs in more detail.

7.3.2. CINEMATIC SCENE DESCRIPTION SURVEY

As the basis for the work reported in this chapter, a HIT template was designed for a survey. With the survey, we wanted to identify suitability criteria for connections between music and narrative videos, described in spontaneous, rich and intuitive ways. Therefore, we asked survey respondents to describe an *idealized cinematic situation, to which to a given music fragment would fit as a soundtrack*. The choice for the cinema genre was made, since cinematic scenes are considered to be a prototypical case in which music soundtracks and visual scenes are connected and jointly create new connotative meaning. Besides, cinematic scenes are works of fiction, and thus allow for unlimited creative freedom in their conception. In order to help respondents in elaborating on their ideas, and facilitate later analysis, the solicited description was broken down into separate free-form text questions related to the scene, its characters and its setting. The full breakdown of all questions of the survey, together with the examples and instructions that accompanied them, are shown in Figure 7.3.

While our description survey has similarities to the earlier cited work of Nattiez [Nattiez, 1973], the types of solicited descriptions differ fundamentally. In the case of Nattiez, the solicited description can be considered a direct visualization or explanation of the music. In our case, we assume the opposite case: the music explains the described cinematic situation. In terms of a causal relationship, the description is the cause and the music the effect, which is common practice in the musical scoring of cinematic productions [Prendergast, 1992].

7.3.3. MUSIC RANKING TASK

Ultimately, we want to place the work of this chapter in a multimedia Music-IR context. Because of this, it is important to not just solicit descriptions for music fragments, but to also verify if music fragments can be realistically recognized and retrieved based on the same type of descriptions. To this end, next to the previously described survey, we designed a HIT in which a cinematic scene description that had been previously supplied would be given to a worker. The worker would then be asked to rank and rate 3 given

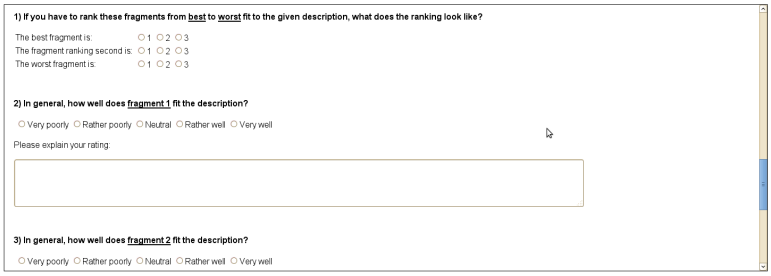


Figure 7.4: Excerpt from the music ranking HIT template.

music fragments according to their fit to the given description. A screenshot of part of the corresponding HIT template is given in Figure 7.4.

The description was presented in the original breakdown form as presented in the previous subsection, e.g separating key character and location descriptions from overall scene descriptions. We only omitted the field in which the original describer explained why he chose the particular scene that was described.

After the given cinematic description, the worker was provided with three music fragments. These fragments were a randomly ordered triplet, out of which one song was the music fragment that had originally triggered the description (from now on indicated as the 'stimulus fragment') . The other two fragments were randomly chosen without replacement from a large royalty-free production music database, which will be described in more detail in Section 7.3.4. The task of the worker was to rank the three fragments according to their fit to the given description. This was done both through explicit indication of the rank positions, as well as a through ratings of the perceived fit of each fragment on a 5-point scale. Finally, the worker had to give an explanation for the given ratings. The input has some redundancy: from the 5-point ratings, a ranking can be devised already. However, this redundancy was conscious and served to ensure answer robustness.

7.3.4. DATA

A music dataset was created with 1086 songs from three royalty-free production music websites[3]. Royalty-free music has the advantage of being legally playable on the Web and reusable; production music has the advantage of having explicitly been written to be used in multimedia contexts.

To prevent listener fatigue, we only consider dataset items with durations between 30 seconds and 5 minutes for our task runs. We chose not to further standardize the fragments lengths, since this would require alterations modifying the original musical discourse. The exact number of songs used differed per task, and will be explained in the following subsection.

[3] `http://incompetech.com` (Kevin MacLeod), `http://www.danosongs.com` (Dan-O) and `http://derekaudette.ottawaarts.com/index2.php` (Derek R. Audette), accessed December 13, 2012.

7.3.5. TASK RUN SPECIFICATIONS

With the tasks described in Section 7.3.2 and 7.3.3 as basis, we performed several batches of experimental runs: (1) an initial description survey run, (2) a description HIT run yielding qualifications for follow-up HITS, (3) a follow-up description HIT run, and (4) a follow-up music ranking HIT run.

INITIAL SURVEY RUN

The first released version of our description survey was more extensive than our outline of Section 7.3.2: apart from a description of a cinematic situation, respondents also had to describe a real-life situation they once experienced to which the music fragment would fit. Furthermore, basic demographic information was asked (gender, age, country, level of (music) education and frequencies of listening to music, watching web videos and watching movies). Finally, since the solicited real-life situation description could be privacy-sensitive, respondents had to explicitly permit the reuse and quoting of their responses.

As the survey was expected to require a much longer completion time (over 15 minutes) than a regular AMT HIT, we at first were concerned about releasing it as such, and instead put up a HIT batch in the AMT Sandbox environment. This would allow us to benefit from the AMT infrastructure, but to handle the recruitment of respondents ourselves. A call for survey participation was disseminated intensively: through a snowball procedure, professional and personal contacts were written and asked to propagate the message by forwarding the message to their own (inter)national professional and personal contacts. Next to this, paper flyers and announcements were distributed on-campus in The Netherlands, and announcements were placed on 6 different web forum communities in which role-playing was strongly featured. Through this, a large and demographically diverse audience was quickly reached.

While our dissemination strategy reached hundreds of people, the actual uptake was not as large as we expected. Multiple contacts forwarded our call, but did not feel obliged or eligible to complete a survey themselves. In addition, unfamiliarity with the AMT platform made several potential respondents reluctant to set up a login account. Upon receiving this feedback, we offered interested people the possibility of using a lab-created account which would be fully independent of their personal information.

For the task, we put up surveys for 200 music fragments meeting the specifications given in Section 7.3.4, with 3 assignments per fragment. After 20 days, 65 completed surveys had been received, from respondents from 13 different countries, on 57 unique music fragments[4]. We had reuse permission of input for 55 of these fragments.

Judging from the responses, respondents found it very hard to envision and describe real-life situation descriptions, but rarely had problems to do this for cinematic scenes. For this reason, in follow-up runs of the survey we did not focus on the real-life situation anymore.

QUALIFICATION RUN

Following the experiences above, we moved to the regular AMT platform after all. At first, we piloted a HIT following the specifications of Section 7.3.2; in addition, the same

[4] Upon the release of a HIT batch, the AMT platform randomizes the order in which assignments are presented to workers, regardless of how many assignments already have been completed for a given HIT.

demographic information as described in Section 7.3.5 was asked. Any worker with a HIT Approval rate of at least 0.9 could perform the task. The reward for successfully completed HITs was $0.09.

We first piloted a HIT task for one song ('Exciting Trailer' by Kevin MacLeod, which was not used in the initial survey), intended to run continuously for several days to harvest as many responses as possible. Therefore, 150 assignments were released for the HIT. Within a few days, it became clear the uptake was much larger than in the case of our initial survey run (more information on this will follow in Section 7.4). Several workers indicated they had not seen a task like ours before on AMT, but despite the longer required working time and the creative load of the task they indicated to enjoy doing the task. Following this enthusiastic response, we released similar HITs for 2 more songs that were not used in our initial survey: 'Mer Bleue Boogie' by Derek R. Audette and 'Origo Vitae' by Dan-O. Once again, these had 150 assignments each, and were intended as opportunities for harvesting many responses for the same song.

Any completed HIT for which the free-form text responses were at a sufficient level of English, and for which there were indications that the music fragment had a relation to the given scene description (as opposed to answers in which workers just described their favorite actors or movie scenes), would grant the corresponding worker a qualification for a follow-up HIT batch: either another description batch as detailed in Section 7.3.5, or a batch of the ranking task detailed in Section 7.3.5.

In total, qualifications were granted to 158 workers. Upon earning a qualification, a worker was e-mailed with a notification. Workers were distributed over batches such that the demographics distributions in all batches would be similar. Qualifications were granted such that a worker would not be able to work on more than one batch at the same time. If the batch to which a worker was assigned finished without the worker having worked on it, he was reassigned to another open batch and notified.

DESCRIPTION TASK

The qualification-requiring description HIT followed Section 7.3.2 and Figure 7.3 exactly. Workers received $0.19 per succesfully completed HIT. Input data consisted of the 55 fragments which were described in our initial survey run, and for which reuse and quoting permission has been granted. Due to a technical issue, the results for one music fragment had to be invalidated after finalization of the batch, causing 54 remaining music fragments to be considered in our further analysis in Section 7.6. Per HIT, 3 assignments were released. However, some songs would ultimately get 4 different descriptions out of this HIT, due to the re-release of a small number of assignments to a motivated worker who had an audio player incompatibility problem.

RANKING TASK

The music ranking HITs exactly followed the specification given in Section 7.3.5. Once again, workers received $0.19 per succesfully completed HIT. For the cinematic scene descriptions for which a worker should rank 3 music fragments according to their fit, we used descriptions for the 55 fragments for which reuse and quoting permission has been granted in our initial survey run. Due to the same technical issue as mentioned in Section 7.3.5, the results for one music fragment had to be invalidated after finaliza-

tion of the batch, causing 54 remaining music fragments to be considered in our further analysis in Section 7.5.

In comparison to the original survey answers, a few minor clean-up changes were made. Firstly, answers that had been given in Dutch (which was allowed for our initial survey, for which recruiting started in The Netherlands) were translated to English. Secondly, basic spelling correction was performed on the answers. Finally, we removed concrete music timestamp references from descriptions (*"The beginning (0-33s): Big overview shots of the protagonist"*), to avoid the HIT becoming a timestamp matching task.

The 2 other fragments to offer with the stimulus fragment were randomly chosen without replacement from our production music database, only considering fragments with a duration between 30 seconds and 5 minutes that had not been used already in our other HIT batches. We released this HIT in 3 separate batches, each requiring a different qualification in order to keep the worker pools mutually exclusive, with 3 assignments per HIT. This choice for multiple batches had two reasons: firstly, having multiple isolated batches ensures a larger minimum number of required workers (in this case, it meant that 3×3 workers were needed per provided description). Secondly, since we generated different triplets for each batch, a stimulus fragment would ultimately be compared against $3 \times 2 = 6$ random fragments.

Table 7.1: Statistics for the initial and qualification survey tasks.

	Running Time	# Workers	Average Worker Age	# Approved Assignments
Initial Run	20d	49	32.6	65
Qual. 'Exciting Trailer'	15d	133	27.8	130
Qual. 'Mer Bleue Boogie'	11d	58	26.7	55
Qual. 'Origo Vitae'	11d	65	26.6	63

	# Rejected Assignments	# Qualifications Granted	Median Completion Time (approved only)
Initial Run	23	N/A	19m 48s
Qual. 'Exciting Trailer'	3	108	9m 16s
Qual. 'Mer Bleue Boogie'	3	40	9m 54 s
Qual. 'Origo Vitae'	2	47	7m 52s

7.4. CROWDSOURCING STATISTICS

General crowdsourcing statistics for all previously described task runs are shown in Tables 7.1 and 7.2. As can be seen in these tables, for almost every task run a few HITs were rejected. For our description tasks (initial survey, qualification task, and qualification-requiring description task), we rejected assignments with blank or nonsense responses in the free-form text fields. For the music ranking batches, we rejected assignments with

Table 7.2: Statistics for the qualification-requiring follow-up tasks.

	Running Time (until completion)	# Qualified Workers	# Actual Workers
Ranking Batch A	10d 15h 10m	43	11 (25.6%)
Ranking Batch B	3d 8h 4m	31	12 (38.7%)
Ranking Batch C	1d 1h 53m	21	12 (57.1%)
Description	10d 5h 5m	89	32 (36.4%)

	# Approved Assignments	# Rejected Assignments	Median Completion Time (approved only)
Ranking Batch A	165	20	7m 33s
Ranking Batch B	165	7	4m 42s
Ranking Batch C	165	0	3m 50s
Description	173	18	8m 23s

blank or nonsense motivations for the chosen ranks and ratings. In all cases, rejected assignments were republished for other workers to complete.

For the batches run on the formal AMT platform, only 3 workers (two in Batch A of the ranking task and one in the description task) were found to challenge the system by consistently copying over a few uninformative answers as assignment responses. In other cases, rejections involved incomplete responses that largely appeared to have resulted from unintentional human error, since the workers causing them typically performed well on other assignments. It is striking to note that the largest absolute number of rejected assignments occurred in our initial survey run. As it turned out, these assignments were performed by a small number of respondents who did not grasp the importance of providing a complete response, leaving all free-form text questions blank because they found them too hard to answer.

Selective worker behavior regarding input data can be noted in Table 7.1 in the number of approved assignments. Out of the 3 possible qualification music fragments, 'Exciting Trailer' received more than twice as many responses as the other two fragments. This cannot be fully explained by the longer running time of the fragment's HIT, and may have been due to a general preference for this fragment.

From Table 7.2, insight can be obtained in the return rate of qualified workers. As can be seen, while many workers indicated they very much liked their qualification HIT, this did not guarantee that they also returned to do more tasks. Our found percentages of qualified workers that returned to perform more HITs are consistent with earlier reported results in [Soleymani and Larson, 2010] and underline the need to accommodate for a substantially larger pool of potential workers, if a certain minimum worker pool size should be met in order for a HIT batch to complete successfully.

Finally, information on effort distribution within batches is shown in Figure 7.5 and Figure 7.6. For all tasks performed on the regular AMT platform, it can be noted that most

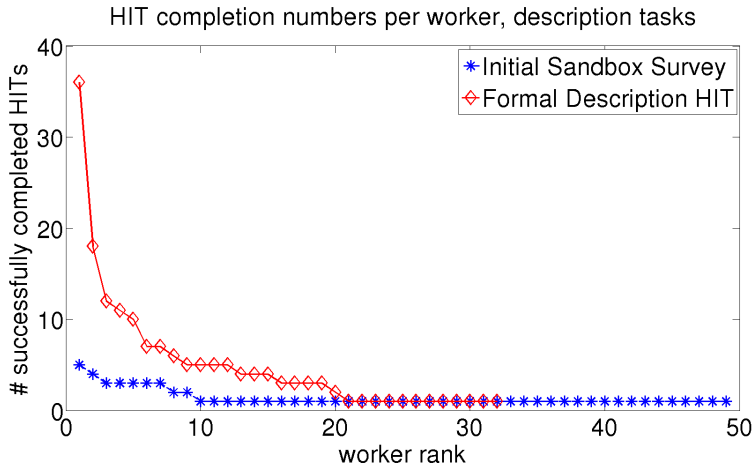


Figure 7.5: HIT effort distribution for the original survey and qualification-requiring description tasks. The horizontal axis represents the rank of workers, when sorted by the number of successfully completed HITs.

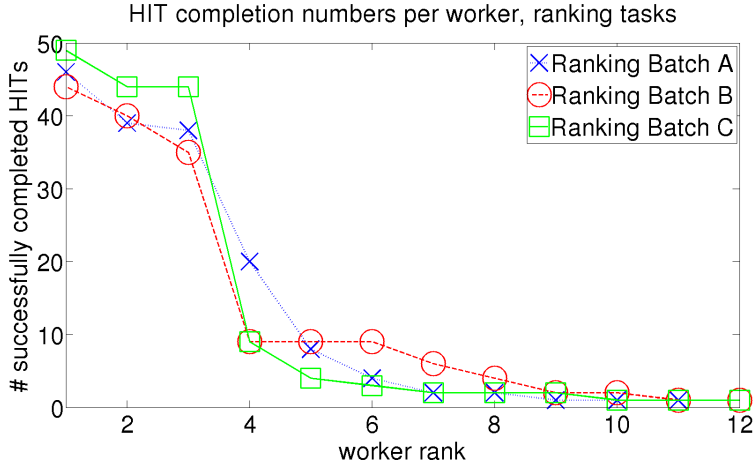


Figure 7.6: HIT effort distribution for the 3 ranking task batches. The horizontal axis represents the rank of workers, when sorted by the number of successfully completed HITs.

of the work is performed by a relatively small amount of motivated workers. This effect is the strongest for the music ranking batches, which all had three highly active workers performing the majority of available assignment tasks (although no single worker managed to perform HITs for all the data in the batch). The qualification-requiring description task required longer completion times and more personal, creative input, which makes it harder for a worker to build up an efficient routine for this task. Indeed, the effort distribution curve for this task as shown in Figure 7.5 is much smoother than in the case of the music ranking batches shown in Figure 7.6.

7.5. MUSIC RANKING TASK RESULTS

In order to test the recognizability of stimulus fragments given a cinematic description, our results analysis will address two questions:

1. To what extent are the provided rankings and ratings consistent across workers?
2. Is the perceived fit to the description larger for the stimulus fragment in comparison to random fragments?

7.5.1. RATING CONSISTENCY

As a measure for inter-rater consistency in case of more than two raters per item, the Fleiss kappa [Fleiss, 1971] is frequently employed (in the Music-IR field this was e.g. done in [Jones et al., 2007]). However, in order to give a statistically valid result, the measure assumes that different items are rated by different sets of raters. For the AMT crowdsourcing setup, this assumption will not hold, since one worker may work on as many assignments as the number of HITs, thus causing large and unpredictable variation in the rater distribution over different items. This variation also makes it difficult to reasonably normalize 5-point scale rating scores per worker to prevent polarization biases. In our analysis, we therefore will treat the given ratings not just as absolute numbers, but also as indicators of relative orderings.

In order to get a measure of rating consistency, we chose to take a similar approach to Urbano et al. [Urbano et al., 2010], in which crowdsourced worker inter-agreement was measured based on preference judgments for melody similarity. In preference judgments HITs, a worker was provided with a melody and two variations. The worker then had to indicate which of the two variations was more similar to the given melody. If n workers are assigned per HIT, $\binom{n}{2}$ worker pairs can be chosen out of this worker pool.

For each pair, an agreement score was computed. If both workers prefer the same item, 2 points are added to an agreement score. If one of the workers indicated that both variations were equally (dis)similar to the melody, 1 point is added to an agreement score. If the workers prefer different items, the agreement score is not increased. The agreement scores for all individual HITs are summed, and divided by the maximum obtainable score of $\binom{n}{2} \times 2 \times$ # HITs.

In our case, for every assignment in every HIT we converted the 3 fragment ratings of a worker into an analoguous form, by transforming them into $\binom{3}{2}$ pairs of orderings. For example, if 3 fragment ratings are $[5;2;3]$, we now encode them as $[5 > 2;\ 5 > 3;\ 2 < 3]$, reflecting their relative ordering rather than absolute rating values, which from a retrieval perspective is reasonable. These are then turned into agreement scores by comparing how pairs of workers judged the relative ordering, in the same way as described above.

Out of the 3 batches × 54 fragments × 3 assignments × 3 ratings per assignment = 1458 ratings in total, 15 ratings were missing in our data. Out of these, for 12 missing ratings (0.82% of the total), the workers had given sufficient information via other input fields to clarify their view on the fragments, showing the benefit of our redundancy mechanisms. In 3 cases (0.21%), a rating was missing due to an audio loading issue.

Since our chosen technique to compute agreement scores does not allow for such missing values, we applied a data imputation procedure before performing the computation. If the unrated fragment was indicated as the best fit, it got the rating score of

Table 7.3: Rating agreement score and rank attributions (in %) for the stimulus fragments.

	Rating Agreement	Rank 1	Rank 2	Rank 3
Batch A	0.7346	83.33	10.49	6.17
Batch B	0.6934	69.14	22.22	8.64
Batch C	0.6317	72.84	14.81	12.35
Overall	0.6866	75.10	15.84	9.05

the second ranked item, regardless if the worker indicated the first item to be a much better fit. Similarly, if the unrated fragment was indicated as the second best fit, it got the rating score of the worst ranked item. If the unrated fragment was indicated as the worst fit, it got the rating score of the second ranked item, regardless if the worker indicated the worst item to be much worse than the second ranked item. In case there was an audio problem, the rating for the corresponding item would be -1, causing the item to automatically be ranked last. This is a conservative approach, such that the results obtained on data to which this policy is applied can be considered to be lower bounds to the actual results. The agreement scores obtained for the three batches are given in the first column Table 7.3. As can be seen, the scores indicate high agreement between the workers regarding the ordering of fragments.

7.5.2. STIMULUS FRAGMENT VS. RANDOM FRAGMENTS

The agreement measure across rating orderings gives only an indication of intra-rater ranking ordering behavior. In order to verify whether the stimulus fragment actually is being ranked higher than the random fragments it occurs with, we consider the frequencies of rank occurrences for the stimulus fragment. As shown in Table 7.3, in a large majority of cases, the stimulus fragment is indeed ranked first for all three batches. Looking across batches at the nine rankings that are available for each song, there always are at least three workers considering the stimulus fragment as the best fitting fragment to a description.

Finally, to still consider the absolute rating numbers across batches, boxplots for the ratings of the provided music triplets in all three batches are shown in Figure 7.7. These are provided per batch, starting with a boxplot of the stimulus fragment, and followed by a boxplot of the ratings obtained for the randomly chosen songs. Here, we did not apply any data imputation procedures, so missing rating values were omitted. Applying the Kruskal-Wallis test to the ratings with a subsequent Tukey-Kramer Honestly Significant Difference Test, it turns out that at $p < 0.05$, the three stimulus fragment groups from the three random fragment groups ($p \approx 0$). Thus, it can be concluded that stimulus fragments are not rated as similarly as randomly chosen fragments, and as our measures regarding ordering and ranking showed, that the stimulus fragments are clearly considered as better fits to the description than the random fragments.

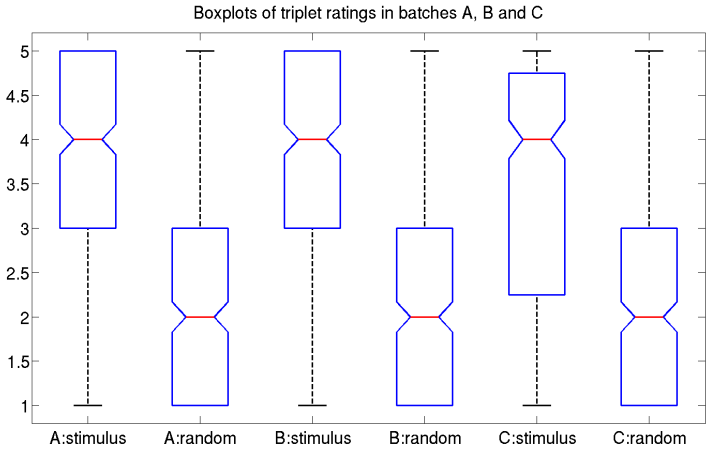


Figure 7.7: Boxplots of the fragment fit ratings (without data imputation on missing values) for the 3 HIT batches. For every batch, ratings are plotted for the original stimulus fragment and the randomly chosen fragments.

7.6. COMMON ELEMENTS: ANALYSIS OF DESCRIPTION RESPONSES

We now arrive at the essential focus of our work: investigating characteristics of the given contextual description responses. Given the amount of individual freedom we granted for these descriptions, analyzing and identifying such characteristics is not straightforward. Nonetheless, in this section we strive to provide deeper insights into them by addressing three questions:

1. Do common narrative elements emerge for a certain song, as soon as we receive a sufficient amount of contextual descriptions for it?

2. What types of common narrative elements can be identified in general for this type of description?

3. Are users capable of indicating why they chose particular scene descriptions for particular songs?

7.6.1. MANY DESCRIPTIONS OF THE SAME MUSIC FRAGMENT

Through our qualification tasks, we acquired a relatively large number of descriptions for several music fragments. Word clouds for obtained scene and location descriptions for the fragments are shown in Figure 7.8 and Figure 7.9. As can be seen, for different fragments, there are different vocabularies and common words. Looking at the full answers the following generalizations and emerging categorizations can be made:

1. 'Exciting Trailer' (108 qualified responses) is a symphonic piece for which the composer stated that *"A militaristic snare drum march begins this piece, reminiscent of*

Eastern Europe during World War II". The fragment gives the majority of worker respondents a sense of *preparation for a situation involving power*. In terms of actors, emerging categories are *fighters* (26.85%) and *heroes* (24.07%). A majority of respondents (64.81%) situates the described scene outdoors, and from the 44 respondents who went as far as indicating a geographic location for the scene, frequently mentioned regions are *Europe* (47.92%) and the *USA* (35.41%).

2. 'Mer Bleue Boogie' (40 qualified responses) is a boogie-woogie piano piece, described by its composer as *"A very up-tempo piano boogie. Heavy swing on this one. It should have your toes tappin'"*. To the worker respondents, the fragment strongly evokes *dancing/party scenes*. In terms of actors, the most frequently mentioned categories are *couples/duos* (37.5%) and *sole male protagonists* (20%). A majority of respondents (65%) situates the described scene indoors. It is striking that the mentioned location categories are very strongly *urban* (91.43%), and out of the 14 respondents who went as far as indicating a geographic location for the scene, 13 (92.86%) situate the scene in the *USA*.

3. 'Origo Vitae' (47 qualified responses) was described by its composer as being *"mysterious and intense"*. To the worker respondents, it evokes mysterious, unknown and sometimes unpleasant situations. The most frequently mentioned actor category is that of *adventurers* (21.28%). A majority of respondents (76.6%) situates the described scene outdoors. In relative terms, this fragment evoked the largest number of geographic area responses (31 respondents), with 15 (49.38%) of the respondents situating the scene in the *Middle East*.

7.6.2. GENERALIZING OVER MORE FRAGMENTS: EVENT STRUCTURE

Already from the responses to our initial survey, we noted apparent agreement between respondents at the level of linguistic event structures. In particular, we identified 4 event structure classes, motivated by the Vendler event typology on events types that can be expressed by human language [Vendler, 1967] (examples in parentheses are taken from the survey responses):

- Class 1: Activity or state with no goal (*"A group of people working in a factory. There is steam everywhere. They work hard, but they are not negative about that."*);
- Class 2: Activity with underspecified goal and no certainty of achieving the goal (*"A scene of awaiting. Something is going to happen."*);
- Class 3: Activity with well-specified goal and no certainty of achieving the goal (*"Warriors are moving fast in the darkness and trying to sneak into their enemies' campsite."*);
- Class 4: Activity with well-specified goal and certainty of achieving the goal (*"Someone is walking in a dense forest and finally arrives at village where the inhabitants are dancing slowly."*)

To verify the validity of these structure classes, we coded all responses belonging to the 54 music fragments that were also used in our ranking tasks, both from the initial survey run and the qualification-requiring description task. In doing this, we took the perspective of the protagonist at the narrative level that Bal [Bal, 2009] calls the *fabula*: the level consisting of the narrative as a pure chain of events, without any 'coloring' of it towards the audience. For 44 out of 54 fragments, a majority of respondents described a story of the same event structure class. For these cases, we studied the amount of agreement by framing the problem as a classification problem, in which the majority vote on the event structure class would be considered as the actual class. By counting actual class occurrences for every majority vote class, we can construct a 'confusion matrix' C, which for our fragments was:

$$C = \begin{bmatrix} 49 & 7 & 12 & 12 \\ 4 & 15 & 3 & 4 \\ 5 & 7 & 43 & 5 \\ 5 & 3 & 0 & 12 \end{bmatrix}$$

where $c_{m,n}$ indicates the number of descriptions for event structure class n, when a majority of respondents described a story of event structure class m.

The most striking result is that class 1 frequently gets mixed with the other classes. This may be because of the way our descriptions were solicited: even when not feeling any goal to be pursued in the music, a respondent could still have been triggered by our questions such that a full story was conceived with a goal in it. However, this situation occurs much less the other way around. If a music fragment invokes a scene with a goal for the majority of respondents, this majority will be large. In addition, respondents are strongly agreeing on uncertainty of reaching a goal (classes 2 and 3).

In order to see how the classes are balanced if many responses are available for a song, we return to the qualification fragments again. The results are shown in Table 7.4. For 'Mer Bleue Boogie', the most frequently occurring class is class 1, but similarly to the confusion matrix, we notice that this is no strong majority either. On the other hand, the other two fragments, in particular 'Exciting Trailer', do have a strong majority class. 'Origo Vitae', which was supposed to be a mystery fragment, has considerably more 'vague goals' (class 2) than the other fragments, or the general trend observed in the confusion matrix.

Table 7.4: Event structure class occurrences (in %) for the qualification fragments.

	1	2	3	4
'Exciting Trailer'	19.62	3.74	57.94	18.69
'Mer Bleue Boogie'	42.5	5.0	37.5	15.0
'Origo Vitae'	8.5	23.4	48.94	19.15

7.6.3. SELF-REPORTED REASONS FOR DESCRIPTIONS

Are workers capable of indicating why the music made them choose their descriptions? This information is relevant to investigate salient high-level description features that

can be used for query construction and to which low-level feature extractors should be mapped.

In order not to bias against songs, we once again considered our set of 231 descriptions for 54 music fragments. We categorized all reported motivations, starting from Tagg's earlier mentioned musical sign topology [Tagg and Clarida, 2003], and adding any other emergent categories. This led to 12 categories:

1. Sonic anaphone: *"The high tones sound like fluttering birds, the lowerer (sic) tones sound like a grumbling character."*
2. Kinetic anaphone: *"The bass line is paced and deliberate. It gave me the sense of someone sneaking around."*
3. Tactile anaphone: *"The music brings to mind picture of something moving slowly, uniformly and languidly like a falling night, setting sun or rising sun, sluggish waters, billowing smoke etc. It also has a touch of romance and warmth. Thus a loving couple in a moonlit night walking slowly in a world of their own came naturally to mind."*
4. Genre synecdoche: *"The saxophone prominence leads me to believe that this film is set in a past generation."*
5. Episodic marker: *"The ascending notes which reach a peak and then decline."*
6. Style indicator: *"The music definitely seemed like something you would hear in a club."*
7. Temporal development: *"It has a beat which (sic) some energy in it, but sometimes it halts, and is there (sic) space for the serious part of a conversation, or just a normal diner (sic)."*
8. Character trait/psychology of actor: *"It is rock enough to make me think of a biker dude."*
9. General quality/atmosphere: *"Because the fragment seems to possess some kind of cool feeling: it is not too fast paced but neither is it too slow. Also the sounds/pads used seem to possess this feeling. "*
10. Seen before: *"Because the fragment reminds me of a robbery in today's popular action movies."*
11. Personal relation: *"It reminds me of my jolly school days."*
12. Intuition: *"Just something that popped in my head the minute I heard this music."*

The answer distribution over these categories is shown in Table 7.5. As it turns out, and was hypothesized at the start of our work, respondents can generally not explain their description reasons at the level of Tagg's musical sign topology (constituted by categories 1 to 6), except in case of the kinetic anaphone category. At the same time, from an application-oriented requirements elicitation perspective, it is good to see that the very individualistic category 11, and the uninformative category 12 only amount to less than 11% of the represented categories. 3 meta-categories stand out:

Table 7.5: Category occurrences (in %) for self-reported scene description reasons.

1	2	3	4	5	6
5.62	10.06	1.12	7.87	0.84	8.15
7	8	9	10	11	12
12.64	7.58	21.63	12.92	5.06	5.62

1. *temporal aspects* (categories 2 and 7, can include 5): temporal developments, movement indicators and episodic changes. In our free-form setting, workers tended to talk much more in terms of characteristics of different episodes that they observed (*"the parts where there are constant harmonies in the background need a more quiet scene than the parts where there is only the beat."*, *"At first, I thought that perhaps a human/hacker type of scene would be appropriate, but it does not explain the breaks in the music."*) than in terms of markers announcing an episode change. imilarly to the findings in [Nattiez, 1973], workers appear to be highly sensitive to temporal developments that indicate a sound change, but from a musical structure analysis or discourse point would not be significant: e.g. *"It has blunt pauses in-between. The end is also blunt and sounds like 'this just does not work'."*, *"The end is a possitive (sic) note and it is as though somebody realises something."*

2. *psychology, quality and atmosphere* (category 9, can include 8): categories relating to Lissa's functions [Lissa, 1965] of indicating a character's psychology, and providing empathy to the viewer. These will be closely related to affective aspects.

3. *previously seen examples* (categories 4, 6 and 10): knowledge of existing films and styles are steering the evocation directions, and workers are aware of this.

7.6.4. FURTHER NOTIONS

There were two more aspects in the received descriptions that we found worth mentioning in this section. While these are currently anecdotical, they seem to conform to theories on cognition, expectation and familiarity as e.g. investigated by Huron [Huron, 2006]. As such, they show interesting opportunities for cross-disciplinary future work.

CULTURAL INFLUENCES

Connotative meaning is largely built on cultural conventions. Overall, our results (which largely reflected Western mass media culture) showed good generalizability measures, indicating that characteristics of this culture are widely recognized, also by workers who are not from countries of this culture. However, there was one example in which culturally specific influences seem present: 'Origo Vitae'. While the Middle East was a popular geographical region, none of the 8 workers from India mentioned this, or any 'exoticism' at all. In music-theoretic terms, the piece features a lowered second degree in comparison to a harmonic major scale, which has grown to be an Arabic cliché in Western art music, but may not be recognized as such by others.

MAJOR VS. MINOR ≠ HAPPY VS. SAD

A common conception regarding the relation of musical keys to affective properties is that major keys imply positive feelings and minor keys negative feelings. In our survey responses, we observed a few situations in which this situation was more nuanced.

For 3 quiet major-key songs, workers agreed on characterizing the song as positive to very positive. However, in the scenes described with it, feelings were mixed: e.g. *"The father dies slowly after having a perfect time with his son. His son gets sad, but at the same time experiences relief due to having the last chance to see him after being apart for 30 years."*, *"The young man in the first scene would be a man with a bright smile, an almost glowing optimism about him, everything in his demeanor a happy and positive. The second scene of this boy would be a older boy, his eyes more wrinkled around the edges, the rings under his eyes deeper, a troubled yet stern look on his face, the face of a man who has stayed up long nights worrying and many times his eyes witnessing the horrible acts war exposes one to, slowly rending tears in this mans happiness and his soul, leaving him a dry and unhappy husk of the man he once was."*.

Furthermore, there were 2 cases of minor-key silent film piano music, which were intended by the original composer as indeed belonging to negatively valenced scenes: *"Short Theme in two tempos for your bad guy"* and *"this staccato piano piece is the tense setting for the classic silent film scene: the heroine is tied to the train tracks, the steam engine train plugging along towards her."*. However, workers do not feel this at all anymore, and rather trigger on the staccato piano playing style, which is considered as bright and uplifting: *"Harry is bulky and brainless. He is always confused. Even if he wants to do the job in a way he feels proud off, he just cant make it. Poor thing."*, *"A few young children are playing hide and seek."*. If any connections are made with existing movies, these are to comedy and slapstick genres (e.g. Laurel and Hardy, The Addams Family).

7.7. CONCLUSION AND OUTLOOK

We presented a study of narrative multimedia descriptions that people connect to music fragments. In order to obtain a sizable number of participants to our study, we employed a crowdsourcing strategy for recruitment. As it turned out, running our highly open-ended task on a mainstream crowdsourcing platform did not present any disadvantages over a more controlled recruiting approach. In contrast, gathering input was more efficient, and there were no significant drawbacks in terms of user input quality.

When provided with earlier obtained descriptions, workers were able to reliably recognize the stimulus fragment that evoked the description. When multiple descriptions were gathered for the same music fragment, fragment-specific profiles emerge in terms of actor, location and story types. Based on the provided worker input, event structure classes that consider the absence or presence of a goal and the certainty of achieving this goal were deduced as a conceptual level at which different free-form worker descriptions will be generalizable. In most of the cases studied, there was a majority preference for an event structure class, and in case the preference considers a narrative involving a goal, this majority is strong. This suggests that automated analysis methods should be conceivable that can map to such goal-oriented classes.

For our current fragment ranking task, our experiments contained a stimulus fragment versus two randomly chosen fragments. This choice to randomly pick the al-

ternative fragments was made to allow for an objective experimental situation, without any potential selection bias. Following the indications regarding salient narrative themes and event structures resulting from our description task, follow-up experiments are imagineable in which the fragment ranking task is repeated, but the alternative fragments are picked in a (semi-)supervised way, considering these findings from our description task.

As we found, workers are not very good at explaining their connotative associations in terms of musical characteristics. Nonetheless, their reports reveal sensitivity to associations between temporal development and musical movement, to affective and psychological effects, and to existing prototypical examples. Two remarks should be made here that need more investigation in the future:

- As for the first self-reported association mechanism of temporal development and movement, the features on which workers trigger seem at a higher level than mere beat or tempo tracking (and related to sound timbre), but at a lower level than formal music structure analysis.

- Future work in this area in the fields of affect and general learning should take into account that connotative associations may involve several signal-unrelated steps (*"I enjoyed this piece of music. It seemed 80s adn (sic) joyful, so that made me think of a romcom's ending. It seemed a little spacy and techy, so that made me think of Geek Romance."*), and that, while consensus is reached on the general affective content of a music fragment, multiple layers of affect will get involved as soon as the multimedia context is involved (*"I thought the music sounded kind of wistful, but on the same hand it also had some sense of pride in it."*).

Our approach to the problem took a cross-disciplinary view inspired by current thought in linguistics and musicology. For several decades, the field of musicology has already been acknowledging and studying the importance of connotative meaning, albeit not from a data-oriented perspective, avoiding any absolutist and positivist approaches. With the rise of the Social Web, we are currently able to take a data-oriented perspective that is not necessarily absolutist or positivist, but rather driven by cultural communities and, as we showed in this chapter, connotation-aware.

Our findings that the connotative connections between free-form and spontaneous descriptions of visual narrative and musical information are strong enough to be recognizable and generalizable opens up new perspectives for multimedia-oriented music querying and retrieval. Music queries in this context do not need to be confined to musical vocabulary, but can be constructed in a user-friendly narrative form created under consideration of the envisioned end result. Such queries constitute versatile and sophisticated multimedia messages, and, as our results suggest, the goal of deepening out the connotative layer underlying these messages is feasible to pursue in the near future.

(a) 'Exciting Trailer'

(b) 'Mer Bleue Boogie'

(c) 'Origo Vitae'

Figure 7.8: Word clouds for qualification fragment scene descriptions by qualified workers. Common English stop words were removed. In addition, the frequently occurring word 'scene' was removed from the word clouds, since it never occurred as an element of the actual scene.

(a) ‘Exciting Trailer’

(b) ‘Mer Bleue Boogie’

(c) ‘Origo Vitae’

Figure 7.9: Word clouds for qualification fragment location descriptions by qualified workers. Common English stop words were removed.

CONCLUSIONS

How far did we get in addressing the ambitions outlined at the beginning of this thesis? In this final chapter of the thesis, we revisit our original goals and ambitions, summarize the main contributions presented in the thesis, discuss further advances caused by work carried out in the context of this thesis, and reflect on outstanding challenges in the field.

In the Introduction of this thesis, we made a strong case for multifaceted approaches to Music-IR. Several key aspects of a multifaceted approach were mentioned:

> *"...it would make sense to consider digital music data as a multimodal phenomenon constituted by hybrid content, which allows various ways of interpretation, and plays an important role in the consumer experience when embedded in various use contexts. [...] In order to fully exploit the potential of digital music information access and consumption (and ease adoption beyond the technical domain), it would be useful to adopt a multidisciplinary (or even more strongly, an inter- or transdisciplinary) strategy, integrating insights and viewpoints from these various research fields throughout the creation process of sophisticated Music-IR technologies."*

While the work in this thesis led to advances in the field along these aspects, in parallel, further related progress has been made in the field thanks to the European PHENICX project. This project was acquired during the progress towards this thesis, and its agenda was strongly inspired by the scene-setting studies of Part I.

Therefore, to emphasize the impact of the paradigms proposed in this thesis on the Music-IR community, in this chapter we will jointly discuss advances which directly were achieved through work reported in this thesis, and related advances in the PHENICX project in which the author of this thesis was involved. To this end, the chapter will start with a short introduction of the PHENICX project. After this, we will discuss achievements and open challenges considering several aspects of a multifaceted approach: considering music as a multimodal phenomenon, allowing various ways of interpretation, enabling novel consumer experiences, embedding music in various use contexts, and adopting multidisciplinarity.

THE PHENICX PROJECT

The European FP7 PHENICX project focuses on investigating the use of technology to create enriched experiences of classical music concerts, which are capable of engaging broader audiences before, during and after concerts take place [Liem et al., 2015]. To do this, concerts are transformed into digital artefacts which are *multimodal*, *multilayer* (i.e. considering multiple layers of relevant information at the same time) and *multi-perspective* (i.e. considering different physical viewpoints from which concerts can be

recorded and analyzed, and different possible user viewpoints in reception depending on what audience segment is considered). On one hand, the project requires the development of novel and improved technologies to handle the specific challenges posed by symphonic classical music data. At the same time, creating a user experience which truly can be enriching rather than distracting is an equally important goal of the project. This requires intensive collaboration between academics, relevant music stakeholders and partners specialized in end-user facing prototype implementation. In this, user factors and attitudes towards classical concert attendance should be carefully considered, which means that real users should be involved throughout the process.

To the best of our knowledge, PHENICX has been the first project to explore these directions in a comprehensive way. It connects multiple relevant academic parties with expertise on necessary Music-IR technologies to important music stakeholders (the Royal Concertgebouw Orchestra in Amsterdam and the ESMUC conservatoire in Barcelona), as well as an integration partner capable of translating academic advances into user-facing products (Video Dock BV). This combination covers a considerable part of the value chain towards the production of digital audiovisually enriched concert experiences. Furthermore, through the music stakeholders, access to realistic user audiences is available.

As a consequence, integrated user-facing prototypes are under development as mentioned in [Liem et al., 2015, Melenhorst et al., 2015], which aim to maximize experience enhancement through the intensive adoption of user-centered design methods [Melenhorst and Liem, 2015]. The design methods explicitly take different audience segments into account: outsiders to classical music, 'casual consumers' who like the genre but do not actively act upon this, and 'heavy consumers' who are established and experienced classical music fans. This way, the project ventures beyond the traditional audiences targeted in classical music marketing, and explores novel ways for addressing potential classical music consumers.

CONSIDERING MUSIC AS A MULTIMODAL PHENOMENON

The concept of treating music as a **multimodal phenomenon constituted by hybrid content** was reflected throughout the thesis.

In Part II ('Data-driven analyses of multiple recorded music performances'), the proposed approaches may not seem multimodal, since they largely considered the audio domain. However, the rationale behind the proposed approaches fundamentally was built upon the viewpoint that music information is established through the hybrid interplay of information from notated scores, human performer interpretation, and resulting auditory renditions, as was defined in Chapter 1.

Part III ('Soundtrack suggestion for user-generated video') of the thesis explicitly considered a multimedia challenge, in which multimodal approaches were necessary to address the problem. The MuseSync system proposed in Chapter 6 therefore combined audio, video and text analysis. In Chapter 7, further connection points were sought between narrative descriptions and musical content.

In parallel, within the PHENICX project, major steps were performed in investigating the use of multimodal information on music performance. Initial work on movement analysis of musicians [Liem et al., 2013] led to further exploration of playing/non-playing

identification of individual musicians in an orchestra based on video analysis [Bazzica et al., accepted for publication]. The use of such information is considered to have support potential in the context of informed source separation and score analysis [Bazzica et al., 2014], and has been of major interest to parties involved in recording and producing multi-camera video recordings of orchestral music.

ALLOWING VARIOUS WAYS OF INTERPRETATION

As mentioned in the general Introduction, the aspect of flexibility towards various ways of interpretation has most strongly been reflected in the work in this thesis. In Part II, various ways of interpreting a music piece were acknowledged as relevant for archive exploration scenarios. Methods were proposed to focus both on similarities and dissimilarities between interpretations, such that 'room for artistic individuality' over the timeline of a piece could be visualized. Next to this, main factors constituting differences in higher-level sound properties of orchestral interpretations could be visualized, and differences between a corpus of various interpreters could be selected and shown.

In Part III, various ways of interpretation were also considered in the context of soundtrack suggestion for user-generated video—more specifically, various intended messages and imposed narratives from users regarding user-generated video. The possibility to allow differing imposed narratives on the same video material was an essential feature of the proposed MuseSync system, which would be overlooked in approaches based on direct video-to-audio signal matching.

In the PHENICX project, being able to compare different musical performances and learn about their differences is a concrete use case within the project. The work in Chapter 5 of this thesis comes forth from this use case. Because of the PHENICX focus on orchestral repertoire, performance analysis had to abandon the traditional focus on piano playing. In doing this, new challenges were posed regarding performance-related parameters that are relevant to consider, while realistic to measure. As a consequence, we proposed a top-down rather than a bottom-up approach to exploratory performance analysis, considering spectrogram analysis to reveal differences in higher-level sound properties between orchestras, such as ensemble balance and timbral coloring.

ENABLING NOVEL CONSUMER EXPERIENCES

Enabling novel consumer experiences was a steering motivation behind much of the work of this thesis. Throughout Part II, the possibility to offer new and enhanced ways to better understand similarities and dissimilarities in archives of multiple interpretations should lead to enhanced and better-informed user explorations of the archives. Regarding Part III, the flexible novel querying mechanism of MuseSync, which does not require the explicit entering of musical search terms, should lead to more natural narrative-based querying of music search engines.

At the same time, it should realistically be noted that in the chapters directly reported in this thesis, the reported work could not yet be validated in the context of end-to-end user-facing consumer experiences. This was due to the high novelty of both the work and problem fields of interest, such that the reported work could not yet reach the maturity and momentum to be embedded in such an end-to-end experience within the scope of

the thesis project.

However, within the PHENICX project, major advances have been made towards assessing novel consumer experiences for classical music, and setting up infrastructure for end-user experience development and verification. With strong focus on orchestral concert event experiences, the project created significant momentum and experience in carrying out user-facing experiments and setting up user-facing prototypes. By now, this has created a better environment and stronger stakeholder network to investigate experience impact of the proposed methods of this thesis in more detail in future work.

Venturing into these directions is by no means trivial, and needs careful consideration at the interface side. As mentioned before, being able to compare performances of the same musical piece is a concrete PHENICX use case, which especially is considered by heavy consumers of classical music to be of interest. At the same time, initial efforts to visualize comparative performance information in an integrated user interface (see Figure 7.10), using performance parameters as earlier proposed by PHENICX project partners [Dixon et al., 2002], have not yet been successful and were deemed in focus group studies to be too technical and too complex to comprehend. Therefore, further investigation is needed in what constitutes understandable ways of explaining differences between performances to audiences, and how these can be designed in such a way that the concert experience is not over-rationalized.

EMBEDDING MUSIC IN VARIOUS USE CONTEXTS

The work in this thesis has been looking into better contextualization of music content. In Part II, recordings were put in the context of a collection of related recordings and subsequently analyzed from a collective top-down viewpoint. This was a novel approach to performance analysis. In the context of ever-expanding digital music archives faced with ad hoc user interests in differing music pieces in the archive, our approach offers a lighter-weight and less annotation-intensive approach than traditional bottom-up approaches to music performance analysis. As such, more scalable and more flexible handling of information in such music archives can be enabled.

The possibility to contextualize content in relation to various envisioned uses was also intensively explored in Part III. Closely relating to the point of 'various ways of interpretation', the methods put forth in this part of the thesis can contextualize user-generated video content in different ways. Furthermore, the insights from Chapter 7 offer valuable insights into a more abstract semantic level at which music can be related to narrative: the level of linguistic event structure.

As mentioned in its introductory section, the PHENICX project explicitly targets different audience segments with various degrees of music sophistication and classical music experience. Especially for audiences which are less initiated in classical music, new ways of communicating about music are investigated. This has led to interest in personalization opportunities and novel ways and connections to relate to music, e.g. leading to the investigation of significance of music events over a timeline as indicated through social feedback [Yadati et al., 2014].

The work in Part III can be considered as relevant to these directions too: cinematic soundtrack music is in terms of sound not far from orchestral classical music, and thus can be used as another 'entry point' to classical music which broader audiences are fa-

Figure 7.10: Initial mockup impression of a user interface explaining performance differences between orchestra recordings, used in PHENICX focus groups targeted at interaction design assessment for the integrated prototype mentioned in [Liem et al., 2015]. While focus group audiences expressed interest in learning about differences between multiple renditions of the same piece, this design was considered as too technical and too complex to comprehend by focus group audiences, requiring further research into more proper ways of displaying this information in an understandable way.

miliar with. Furthermore, as mentioned in Chapter 6, the use of collaborative web resources has potential to reveal culturally specific associations to music, which again can be used to contextualize music and provide recognizable 'anchoring' points of understanding to new audiences in better ways.

ADOPTING MULTIDISCIPLINARITY

Throughout this thesis, efforts to employ multidisciplinary methods and pinpoint ways that can stimulate viable multidisciplinary connections were displayed continuously. As for Part II, the interest in performance differences and more subjective and artistic aspects of recordings is of strong interest to music scholars and performers. Proper methods to surface this information thus are helpful to stimulate more intensive and comprehensive digital information access and technology adoption by these audiences (while at the same time, the interfacing challenges mentioned in the context of Figure 7.10 still hold).

In Part III, our approaches in the MuseSync system proposed in Chapter 6 explicitly built on existing thoughts in Musicology, Media Studies and Music Psychology, while still proposing solutions rooted in data analysis and automated mining of large collaborative web resources. Furthermore, our analysis in Chapter 7 of stable elements in user descriptions revealed interesting relations to concepts from Linguistics, thus providing another opportunity for multidisciplinary collaboration.

Generally spoken, in musicological study at the beginning of the 20th century, music was mostly seen as a *positivist* phenomenon. Under this view, music could be studied in an absolute and independent sense, and was considered to be fundamentally represented by scores. However, from the 1980s onwards, a new stream of thinking emerged in which subjectivity, criticism and value judgements was emphasized, in which the contextual and social surroundings of music became important, and in which 'the music itself' even became a taboo concept [Wiering and Volk, 2011]. As such, culture, context and identity have become major topics in modern musicology research. In a certain sense, similar developments occurred in Music-IR [Kaminskas and Ricci, 2012, Knees and Schedl, 2013, Liem et al., 2011b]: apart from studying aspects of the isolated musical object, the roles of context, usage scenarios and relations to other domains and modalities have become increasingly important. Furthermore, the Social Web has increasingly been studied as a potential source for getting information on context and typical user-entered labels of music objects. We believe this opens an important door to collaboration opportunities which truly will be able to interest multiple stakeholder domains at once, including the audiences mentioned in Chapter 2, and that these topics even can form proper crossroads for inter- and transdisciplinary approaches to music retrieval.

Finally, within the PHENICX project, a multidisciplinary consortium with broad value chain coverage is targeting the challenge of creating enriched concert experience. This did not only lead to audience validations in realistic settings, but also to valuable joint work and new insights into technology-supported promotion possibilities for concert events, which already is influencing marketing strategies of music stakeholders.

In conclusion, throughout this thesis, all multifaceted aspects of music which were set out in the Introduction have been covered, and in parallel to the thesis, further

progress along many of these aspects was achieved in the PHENICX project. Thanks to this, the way is paved to continue along all directions of future work proposed in the thesis, in an environment which is becoming increasingly facilitating towards the impact potential assessment of the proposed technologies at the consumer experience side.

BIBLIOGRAPHY

Jakob Abesser, Olivier Lartillot, Christian Dittmar, Tuomas Eerola, and Gerald Schuller. Modeling musical attributes to characterize ensemble recordings using rhythmic audio features. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 189 – 192, Prague, Czech Republic, May 2011.

Jean-Julien Aucouturier. Sounds like Teen Spirit: Computational Insights into the Grounding of Everyday Musical Terms. In James W. Minett and William S-Y. Wang, editors, *Language, Evolution and the Brain*. Academia Sinica Press, 2009.

Jean-Julien Aucouturier and François Pachet. Improving Timbre Similarity: How high is the sky? *Journal of negative results in speech and audio sciences*, 1(1), 2004.

Jean-Julien Aucouturier, François Pachet, Pierre Roy, and Anthony Beurivé. Signal + Context = Better Classification. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 425–430, Vienna, Austria, 2007.

David Bainbridge, Brook J. Novak, and Sally Jo Cunningham. A user-centered design of a personal digital library for music exploration. In *Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 149–158, Surfer's Paradise, Australia, 2010.

Mieke Bal. *Narratology — Introduction to the Theory of Narrative*. University of Toronto Press, 3rd edition, 2009.

Luke Barrington, Reid Oda, and Gert Lanckriet. Smarter than Genius? Human Evaluation of Music Recommender Systems. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 357–362, Kobe, Japan, October 2009a.

Luke Barrington, Douglas Turnbull, Mehrdad Yazdani, and Gert Lanckriet. Combining Audio Content and Social Context for Semantic Music Discovery. In *Proceedings of the 32nd International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 387–394, Boston, USA, 2009b.

Roland Barthes. *Image Music Text*. Hill and Wang, 1977.

Mathieu Barthet and Simon Dixon. Ethnographic Observations of Musicologists at the British Library: Implications for Music Information Retrieval. In *Proceedings of the 12th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 353–358, Miami, USA, 2011.

Alessio Bazzica. Generazione automatica di video musicali automatic music video generation. Master's thesis, Università degli studi di Firenze, 2012.

Alessio Bazzica, Cynthia C. S. Liem, and Alan Hanjalic. Exploiting Instrument-wise Playing/Non-Playing Labels for Score Synchronization of Symphonic Music. In *Proceedings of the 15th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 201–206, Taipei, Taiwan, 2014.

Alessio Bazzica, Cynthia C. S. Liem, and Alan Hanjalic. On Detecting the Playing/Non-Playing Activity of Musicians in Symphonic Music Videos. *Computer Vision and Image Understanding*, accepted for publication.

Nicholas J. Belkin. Some(what) Grand Challenges for Information Retrieval. *SIGIR Forum*, 42(1), June 2008.

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

Sergio Benini, Luca Canini, and Riccardo Leonardi. A Connotative Space for Supporting Movie Affective Recommendation. *IEEE Transactions on Multimedia*, 13(6):1365–1370, December 2011.

Thierry Bertin-Mahieux, Douglas Eck, and Michael Mandel. Automatic Tagging of Audio: The State-of-the-Art. In Wenwu Wang, editor, *Machine Audition: Principles, Algorithms and Systems*, pages 334–352. IGI Publishing, 2010.

Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Zapata, and Xavier Serra. ESSENTIA: an Audio Analysis Library for Music Information Retrieval. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 493–498, Curitiba, Brazil, 2013.

John Bradley. No Job for Techies: Technical contributions to research in the Digital Humanities. In *Digital Humanities*, University of Maryland, July 2009.

Christine D. Brown. Straddling the humanities and social sciences: The research process of music scholars. *Library & Information Science Research*, 24(1):73–94, 2002.

Rui Cai, Chao Zhang, Chong Wang, Lei Zhang, and Wei-Ying Ma. MusicSense: Contextual Music Recommendation using Emotional Allocation Modeling. In *Proceedings of the 15th Annual ACM International Conference on Multimedia (ACM MM)*, pages 553–556, Augsburg, Germany, 2007.

Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma. A Review of Algorithms for Audio Fingerprinting. In *Proceedings of the 5th IEEE Workshop on Multimedia Signal Processing (MMSP)*, pages 169–173, St. Thomas, Virgin Islands, USA, 2002.

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

Michael Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, April 2008b.

Oscar Celma. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer, 2010.

Seymour B. Chatman. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press, 1980.

Eric Cheng and Elaine Chew. Quantitative Analysis of Phrasing Strategies in Expressive Performance: Computational Methods and Analysis of Performances of Unaccompanied Bach for Solo Violin. *Journal of New Music Research*, 37:325–338, December 2008.

Bernard C. K. Choi and Anita W. P. Pak. Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. definitions, objectives, and evidence of effectiveness. *Clinical and investigative medicine*, 29(6):351–64, 2006.

Annabel J. Cohen. How music influences the interpretation of film and video: Approaches from experimental psychology. In Roger A. Kendall and Roger W. Savage, editors, *Selected Reports in Ethnomusicology: Perspectives in Systematic Musicology*, volume 12, pages 15–36. Department of Ethnomusicology, University of California, Los Angeles, 2005.

Carlo Colombo, Alberto Del Bimbo, and Pietro Pala. Retrieval of commercials by semantic content: The semiotic perspective. *Multimedia Tools and Applications*, 13:93–118, 2001.

Arshia Cont. A coupled duration-focused architecture for realtime music to score alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):974–987, June 2010.

Arshia Cont. On the Creative Use of Score Following and Its Impact on Research. In *Proceedings of the 8th Sound and Music Computing Conference (SMC 2011)*, Padova, Italy, July 2011.

Nicholas Cook. *Music — A Very Short Introduction*. Oxford University Press, New York, USA, 1998.

Nicholas Cook. Towards the compleat musicologist? In *Proceedings of the 6th International Symposium for Music Information Retrieval (ISMIR) [invited talk]*, London, UK, 2005.

Nicholas Cook. Performance Analysis and Chopin's Mazurkas. *Musicae Scientiae*, 11(2): 183–207, Fall 2007.

Sally Jo Cunningham, Nina Reeves, and Matthew Britland. An Ethnographic Study of Music Information Seeking: Implications for the Design of a Music Digital Library. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 5–17, Houston, USA, 2003.

David Damm, Christian Fremerey, Verena Thomas, Michael Clausen, Frank Kurth, and Meinard Müller. A Digital Library Framework for Heterogeneous Music Collections—from Document Acquisition to Cross-Modal Interaction. *International Journal on Digital Libraries*, 12:53–71, 2012.

Roger B. Dannenberg. An On-Line Algorithm for Real-Time Accompaniment. In *Proceedings of the International Computer Music Conference*, pages 193–198, Paris, France, 1984.

Roger B. Dannenberg and Christopher Raphael. Music Score Alignment and Computer Accompaniment. *Communications of the ACM*, 49(8):38–43, 2006.

Peter Desain and Henkjan Honing. Does expressive timing in music performance scale proportionally with tempo? *Psychological Research*, 56(4):285–292, July 1994.

Johanna Devaney, Michael I. Mandel, Daniel P. W. Ellis, and Ichiro Fujinaga. Automatically extracting performance data from recordings of trained singers. *Psychomusicology: Music, Mind & Brain*, 21:108–136, 2011.

Simon Dixon, Werner Goebl, and Gerhard Widmer. Real time tracking and visualisation of musical expression. In *Music and Artificial Intelligence*, pages 58–68. Springer, 2002.

J. Stephen Downie. Whither MIR Research: Thoughts about the Future. In *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR)*, Bloomington, USA, 2001.

J. Stephen Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science & Technology*, 29(4):247–255, 2008.

J. Stephen Downie and Sally Jo Cunningham. Toward a Theory of Music Information Retrieval Queries: System Design Implications. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, pages 299–300, Paris, France, 2002.

J. Stephen Downie, Donald Byrd, and Tim Crawford. Ten Years of ISMIR: Reflections on Challenges and Opportunities. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 13–18, Kobe, Japan, 2009.

Douglas Eck, Paul Lamere, Thierry Bertin-Mahieux, and Stephen Green. Automatic Generation of Social Tags for Music Recommendation. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2007.

Sebastian Ewert, Meinard Müller, and Peter Grosche. High Resolution Audio Synchronization Using Chroma Onset Features. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1869 – 1872, Taipei, Taiwan, 2009.

Jiashi Feng, Bingbing Ni, and Shuicheng Yan. Auto-generation of professional background music for home-made videos. In *Proceedings of the 2nd International Conference on Internet Multimedia Computing and Service (ICIMCS)*, pages 15–18, 2010.

Benjamin Fields. *Contextualize Your Listening: The Playlist as Recommendation Engine*. PhD thesis, Goldsmiths, University of London, 2011.

Joseph L. Fleiss. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Jonathan Foote, Matthew Cooper, and Andreas Girgensohn. Creating music videos using automatic media analysis. In *Proceedings of the 10th ACM International Conference on Multimedia (ACM MM)*, pages 553–560, New York, USA, 2002. ACM Press.

Anders Friberg and Johan Sundberg. Does music performance allude to locomotion? A model of final *ritardandi* derived from measurements of stopping runners. *Journal of the Acoustic Society of America*, 105(3):1469–1484, March 1999.

Hiromasa Fujihara, Masataka Goto, Jun Ogata, and Hiroshi Okuno. LyricSynchronizer: Automatic Synchronization Method Between Musical Audio Signals and Lyrics. *IEEE Journal on Selected Topics in Signal Processing*, 5(6):1252–1261, 2011.

Werner Goebl and Gerhard Widmer. On the use of computational methods for expressive music performance. In Tim Crawford and Lorna Gibson, editors, *Modern Methods for Musicology: Prospects, Proposals and Realities*, chapter 7, pages 93–113. Ashgate, London, UK, 2008.

Werner Goebl, Simon Dixon, Giovanni De Poli, Anders Friberg, Roberto Bresin, and Gerhard Widmer. "Sense" in expressive music performance: Data acquisition, computational studies, and models. In Pietro Polotti and Davide Rocchesso, editors, *Sound to sense, sense to sound: a state of the art in sound and music computing*. Logos Verlag, 2007.

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

Masataka Goto. Music Listening in the Future: Augmented Music-Understanding Interfaces and Crowd Music Listening. In *Proceedings of the 42nd AES Conference on Semantic Audio*, pages 21–30, Ilmenau, Germany, 2011.

Maarten Grachten and Gerhard Widmer. Who is who in the end? Recognizing pianists by their final ritardandi. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 51–56, Kobe, Japan, October 2009.

Maarten Grachten, Martin Gasser, Andreas Arzt, and Gerhard Widmer. Automatic Alignment of Music Performances with Structural Differences. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 607–612, Curitiba, Brazil, 2013.

Donald J. Grout and Claude Palisca. *A History of Western Music*. W. W. Norton & Co, New York, USA, 2000.

Alan Hanjalic, Rainer Lienhart, Wei-Ying Ma, and John R. Smith. The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away? *Proceedings of the IEEE*, 96(4):541–547, April 2008.

Xiao Hu and J. Stephen Downie. When Lyrics outperform Audio for Music Mood Classification: a Feature Analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 619–624, Utrecht, The Netherlands, August 2010a.

Xiao Hu and J. Stephen Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 159–168, Surfer's Paradise, Australia, 2010b.

Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. Optimization-based automated home video editing system. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):572–583, 2004.

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

Charles Inskip, Andrew MacFarlane, and Pauline Rafferty. Music, Movies and Meaning: Communication in Film-makers' Search for Pre-existing Music, and the Implications for Music Information Retrieval. In *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR)*, pages 477–482, Philadelphia, USA, 2008.

Charles Inskip, Andrew MacFarlane, and Pauline Rafferty. Upbeat and Quirky, with a Bit of a Build: Interpretive Repertoires in Creative Music Search. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 655–661, Utrecht, The Netherlands, 2010.

Sylvie Jeannin and Ajay Divakaran. MPEG-7 visual motion descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):720–724, 2001.

Cyril Joder, Slim Essid, and Gaël Richard. A Conditional Random Field Framework for Robust and Scalable Audio-to-Score Matching. *IEEE Transactions on Audio, Speech and Language Processing*, 19(8):2385–2397, November 2011.

M. Cameron Jones, J. Stephen Downie, and Andreas F. Ehmann. Human Similarity Judgments: Implications for the Design of Formal Evaluations. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 539–542, Vienna, Austria, 2007.

Patrik N. Juslin. Five facets of musical expression: a psychologist's perspective on music performance. *Psychology of Music*, 31(3):273–302, July 2003.

Kathryn M. Kalinak. *Settling the score : music and the classical Hollywood film*. University of Wisconsin Press, 1992.

Marius Kaminskas and Francesco Ricci. Contextual music information retrieval: State of the art and challenges. *Computer Science Review*, 6(2–3):89–119, 2012.

Min-Yen Kan, Ye Wang, Denny Iskandar, Tin Lay Nwe, and Arun Shenoy. LyricAlly: Automatic Synchronization of Textual Lyrics to Acoustic Music Signals. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):338–349, 2008.

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 *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 255–266, Utrecht, The Netherlands, August 2010.

Peter Knees and Markus Schedl. A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 10(1), December 2013.

Peter Knees, Tim Pohle, Markus Schedl, and Gerhard Widmer. A Music Search Engine Built upon Audio-based and Web-based Similarity Measures. In *Proceedings of the 30th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 447–454, Amsterdam, The Netherlands, 2007.

Ian Knopke and Frauke Jürgensen. Symbolic Data Mining in Musicology. In Tao Li, Mitsunori Ogihara, and George Tzanetakis, editors, *Music Data Mining*, pages 327–345. CRC Press, Boca Raton, FL, 2011.

Katerina Kosta, Oscar F. Bandtlow, and Elaine Chew. Practical implications of dynamic markings in the score: Is piano always piano? In *Proceedings of the 53rd International AES Conference on Semantic Audio*, London, UK, January 2014.

Fang-Fei Kuo, Meng-Fen Chiang, Man-Kwan Shan, and Suh-Yin Lee. Emotion-based music recommendation by association discovery from film music. In *Proceedings of the 13th Annual ACM International Conference on Multimedia (ACM MM)*, pages 507–510, Singapore, 2005.

Frank Kurth and Meinard Müller. Efficient Index-Based Audio Matching. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):382–395, February 2008.

Frank Kurth, Meinard Müller, Christian Fremerey, Yoon ha Chang, and Michael Clausen. Automated Synchronization of Scanned Sheet Music with Audio Recordings. In *Proceedings of the International Conference on Music Information Retrieval)*, pages 261–266, Vienna, Austria, September 2007.

Edith Lang and George West. *Musical accompaniment of moving pictures — a practical manual for pianists and organists and an exposition of the principles underlying the musical interpretation of moving pictures.* The Boston Music Company, 1920.

Edith Law. Human Computation for Music Classification. In Tao Li, Mitsunori Ogihara, and George Tzanetakis, editors, *Music Data Mining*, pages 281–301. CRC Press, 2011.

Edith Law and Luis Von Ahn. Input-agreement: A new mechanism for collecting data using human computation games. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI) 2009*, pages 1197–1206, Boston, USA, 2009.

Jin Ha Lee. Analysis of user needs and information features in natural language queries seeking user information. *Journal of the American Society for Information Science and Technology (JASIST)*, 61(5):1025–1045, 2010.

Jin Ha Lee and Sally Jo Cunningham. The Impact (or Non-Impact) of User Studies in Music Information Retrieval. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 391–396, Porto, Portugal, 2012.

Jin Ha Lee, Trent Hill, and Lauren Work. What Does Music Mood Mean for Real Users? In *Proceedings of the 2012 iConference*, pages 112–119, Toronto, Canada, 2012.

Cheng-Te Li and Man-Kwan Shan. Emotion-based Impressionism Slideshow with Automatic Music Accompaniment. In *Proceedings of the 15th Annual ACM International Conference on Multimedia (ACM MM)*, pages 839–842, Augsburg, Germany, 2007.

Thomas Lidy and Pieter van der Linden. Report on 3rd CHORUS+ Think-Tank: Think-Tank on the Future of Music Search, Access and Consumption, MIDEM 2011. Technical report, CHORUS+ European Coordination Action on Audiovisual Search, Cannes, France, March 15 2011.

Elad Liebman, Eitan Ornoy, and Benny Chor. A Phylogenetic Approach to Music Performance Analysis. *Journal of New Music Research*, 41:195–222, June 2012.

Cynthia C. S. Liem and Alan Hanjalic. Cover song retrieval: A comparative study of system component choices. In *Proceedings of the 10th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 573–578, Kobe, Japan, October 2009.

Cynthia C. S. Liem and Alan Hanjalic. Expressive Timing from Cross-Performance and Audio-based Alignment Patterns: An Extended Case Study. In *Proceedings of the 12th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 519–524, Miami, USA, 2011.

Cynthia C. S. Liem, Alan Hanjalic, and Craig Stuart Sapp. Expressivity in musical timing in relation to musical structure and interpretation: A cross-performance, audio-based approach. In *Proceedings of the 42nd International AES Conference on Semantic Audio*, pages 255–264, Ilmenau, Germany, July 2011a.

Cynthia C. S. Liem, Meinard Müller, Douglas Eck, George Tzanetakis, and Alan Hanjalic. The Need for Music Information Retrieval with User-Centered and Multimodal Strategies. In *Proceedings of the 1st International Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies (MIRUM) at ACM Multimedia*, pages 1–6, Scottsdale, USA, November 2011b.

Cynthia C. S. Liem, Andreas Rauber, Thomas Lidy, Richard Lewis, Christopher Raphael, Joshua D. Reiss, Tim Crawford, and Alan Hanjalic. Music Information Technology and Professional Stakeholder Audiences: Mind the Adoption Gap. In Meinard Müller, Masataka Goto, and Markus Schedl, editors, *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, pages 227–246. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2012.

Cynthia C. S. Liem, Alessio Bazzica, and Alan Hanjalic. Looking Beyond Sound: Unsupervised Analysis of Musician Videos. In *Proceedings of the 14th International Workshop on Image and Audio Analysis for Multimedia Interactive services (WIA^{2}MIS 2013)*, Paris, France, 2013.

Cynthia C. S. Liem, Emilia Gómez, and Markus Schedl. PHENICX: Innovating the Classical Music Experience. In *Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME) (EU-Projects Program)*, Torino, Italy, June 2015.

Zofia Lissa. *Ästhetik der Filmmusik*. Henschelverlag, Berlin, Germany, 1965.

Beth Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *Proceedings of the International Symposium on Music Information Retrieval*, October 2000.

Michael I. Mandel, Douglas Eck, and Yoshua Bengio. Learning Tags that Vary Within a Song. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 399–404, Utrecht, The Netherlands, August 2010.

Winter Mason and Siddharth Suri. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1):1–23, March 2012.

Matthias Mauch, Hiromasa Fujihara, and Masataka Goto. Integrating Additional Chord Information into HMM-Based Lyrics-to-Audio Alignment. *IEEE Transactions on Audio, Speech and Language Processing*, 20(1):200 –210, 2011.

Rudolf Mayer, Robert Neumayer, and Andreas Rauber. Combination of audio and lyrics features for genre classification in digital audio collections. In *Proceedings of the Annual ACM International Conference on Multimedia (ACM MM)*, pages 159–168, Vancouver, Canada, 2008.

Cory McKay, John Ashley Burgoyne, Jason Hockman, Jordan B. L. Smith, Gabriel Vigliensoni, and Ichiro Fujinaga. Evaluating the genre classification performance of lyrical features relative to audio, symbolic and cultural features. In *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR)*, pages 213–218, Philadelphia, USA, 2008.

Mark S. Melenhorst and Cynthia C. S. Liem. Put the Concert Attendee in the Spotlight. A User-Centered Design and Development approach for Classical Concert Applications. In *Proceedings of the 16th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 800–806, Málaga, Spain, 2015.

Mark S. Melenhorst, Ron van der Sterren, Andreas Arzt, Agustín Martorell, and Cynthia C. S. Liem. A Tablet App to Enrich the Live and Post-Live Experience of Classical Concerts. In *Proceedings of the 1st International Workshop on Interactive Content Consumption (WSICC) at TVX 2015*, Brussels, Belgium, June 2015.

Leonard B. Meyer. *Emotion and meaning in music*. The University of Chicago Press, 1968.

James A. Moorer. Audio in the New Millennium. *Journal of the Audio Engineering Society*, 48:490–498, May 2000.

Meinard Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

Meinard Müller. New Developments in Music Information Retrieval. In *Proceedings of the 42nd AES Conference on Semantic Audio*, pages 11–20, Ilmenau, Germany, 2011.

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.

Meinard Müller, Frank Kurth, and Michael Clausen. Audio matching via chroma-based statistical features. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 288–295, London, UK, 2005.

Meinard Müller, Verena Konz, Andi Scharfstein, Sebastian Ewert, and Michael Clausen. Towards automated extraction of tempo parameters from expressive music recordings. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 69–74, Kobe, Japan, October 2009.

Meinard Müller, Michael Clausen, Verena Konz, Sebastian Ewert, and Christian Fremerey. A Multimodal Way of Experiencing and Exploring Music. *Interdisc. Sci. Rev. (ISR)*, 35(2):138–153, 2010a.

Meinard Müller, Peter Grosche, and Craig Stuart Sapp. What makes beat tracking difficult? a case study on chopin mazurkas. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 649–654, Utrecht, The Netherlands, August 2010b.

Meinard Müller, Daniel P. W. Ellis, Anssi Klapuri, and Gaël Richard. Signal Processing for Music Analysis. *IEEE Journal on Selected Topics in Signal Processing*, 5(6):1088–1110, 2011.

Frank Nack and Lynda Hardman. Denotative and connotative semantics in hypermedia: proposal for a semiotic-aware architecture. *New Review of Hypermedia and Multimedia*, 7(1):7–37, 2001.

Jean-Jacques Nattiez. Y a-t-il une diégèse musicale? In Peter Faltin and Hans-Peter Reinecke, editors, *Musik und Verstehen — Aufsätze zur semiotischen Theorie, Ästhetik und Soziologie der musikalischen Rezeption*, pages 247–257. Arno Volk Verlag, Cologne, Germany, 1973.

Steven R. Ness and George Tzanetakis. SOMba: Multiuser music creation using Self-Organizing Maps and Motion Tracking. In *Proceedings of the International Computer Music Conference*, pages 403–406, Montreal, Canada, 2009.

Steven R. Ness, Thomas C. Walters, and Richard F. Lyon. Auditory Sparse Coding. In Tao Li, Mitsunori Ogihara, and George Tzanetakis, editors, *Music Data Mining*, pages 77–93. CRC Press, 2011.

Nielsen. Digital music consumption and digital music access, January 2011.

Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, pages 557–566, New York, USA, 2010. ACM.

Nicola Orio. Music Retrieval: A Tutorial and Review. *Foundations & Trends in IR*, 1(1): 1–90, 2006.

OUP, 2008. *Concise Oxford English Dictionary*. Oxford University Press, 2008.

Caroline Palmer. Anatomy of a performance: Sources of musical expression. *Music Perception*, 13:433–453, Spring 1996.

Caroline Palmer and Sean Hutchins. What is musical prosody? In Brian H. Ross, editor, *The psychology of learning and motivation*, volume 46, pages 245–278. Elsevier, 2006.

Elias Pampalk, Arthur Flexer, and Gerhard Widmer. Improvements of Audio-based Music Similarity and Genre Classification. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 628–633, London, UK, 2005.

Gabriele Paolacci, Jesse Chandler, and Panos G. Ipeirotis. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419, August 2010.

Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-Based Music Structure Analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 625–636, Utrecht, The Netherlands, August 2010.

Theo Pavlidis. The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away? - An Answer. `http://www.theopavlidis.com/technology/CBIR/summaryB.htm`, accessed September 2011, 2008.

Amandine Penel and Carolyn Drake. Sources of timing variations in music performance: a psychological segmentation model. *Psychological Research*, 61(1):12–32, March 1998.

Enrique Perez Gonzalez and Joshua D. Reiss. Automatic Mixing: Live Downmixing Stereo Panner. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx)*, Bordeaux, France, September 2007.

Enrique Perez Gonzalez and Joshua D. Reiss. An automatic maximum gain normalization technique with applications to audio mixing. In *Proceedings of the 124th AES Convention*, Amsterdam, The Netherlands, May 2008a.

Enrique Perez Gonzalez and Joshua D. Reiss. Determination and correction of individual channel time offsets for signals involved in an audio mixture. In *Proceedings of the 125th AES Convention*, San Francisco, USA, October 2008b.

Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees, and Gerhard Widmer. On Rhythm and General Music Similarity. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 525–530, Kobe, Japan, October 2009.

Roy M. Prendergast. *Film music : a neglected art — a critical study of music in films*. Norton, 1992.

Christopher Raphael. A Probabilistic Expert System for Automatic Musical Accompaniment. *Journal of Computational and Graphical Statistics*, 10(3):487–512, 2001a.

Christopher Raphael. Music Plus One: A System for Expressive and Flexible Musical Accompaniment. In *Proceedings of the International Computer Music Conference (ICMC)*, Havana, Cuba, September 2001b.

Christopher Raphael. Music Plus One and Machine Learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pages 21–28, Haifa, Israel, June 2010.

Bruno Repp. Relational invariance of expressive microstructure across global tempo changes in music performance: An exploratory study. *Psychological Research*, 56(4): 269–284, July 1994.

Bruno Repp. A microcosm of musical expression. I. Quantitative analysis of pianist's timing in the initial measures of Chopin's Etude in E major. *Journal of the Acoustical Society of America*, 104(2):1085–1100, August 1998.

Craig Stuart Sapp. Comparative analysis of multiple musical performances. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 497–500, Vienna, Austria, September 2007.

Craig Stuart Sapp. Hybrid numeric/rank similarity metrics for musical performance analysis. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, pages 501–506, Philadelphia, USA, September 2008.

Markus Schedl. *Automatically Extracting, Analyzing, and Visualizing Information on Music Artists from the World Wide Web*. PhD thesis, Johannes Kepler University, Linz, Austria, June 2008.

Markus Schedl and Peter Knees. Context-based Music Similarity Estimation. In *Proceedings of the 3rd International Workshop on Learning the Semantics of Audio Signals (LSAS 2009)*, Graz, Austria, December 2009.

Markus Schedl, Sebastian Stober, Emilia Gómez, Nicola Orio, and Cynthia C. S. Liem. User-Aware Music Retrieval and Recommendation. In Meinard Müller, Masataka Goto, and Markus Schedl, editors, *Multimodal Music Processing*, Dagstuhl Follow-Ups, pages 135–156. Schloss Dagstuhl - Leibniz Center für Informatik GmbH, 2012.

Erik M. Schmidt and Youngmoo E. Kim. Prediction of Time-Varying Musical Mood Distributions Using Kalman Filtering. In *Proceedings of the Ninth IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 655–660, Washington D.C., USA, 2010.

Björn Schuller and Felix Burkhardt. Learning with Synthesized Speech for Automatic Emotion Recognition. In *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5150–5153, Dallas, USA, 2010.

Carl E. Seashore. *Psychology of music.* University of Iowa Press, Iowa City, 1938.

Joan Serrà, Emilia Gómez, Perfecto Herrera, and Xavier Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech and Language Processing*, 16:1138–1151, August 2008.

Malcolm Slaney. Web-Scale Multimedia Analysis: Does Content Matter? *IEEE MultiMedia*, 18(2):12–15, April 2011.

Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(23):1349–1380, December 2000.

Charles Percy Snow. *The Two Cultures.* Cambridge University Press, 1993.

Mohammad Soleymani and Martha Larson. Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus. In *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, Geneva, Switzerland, 2010.

Mohamed Sordo, Fabien Gouyon, and Luis Sarmento. A Method for Obtaining Semantic Facets of Music Tags. In *Proceedings of the International ACM Workshop on Music Recommendation and Discovery (WOMRAD)*, Barcelona, Spain, 2010.

Sebastian Stober. Adaptive Distance Measures for Exploration and Structuring of Music Collections. In *Proceedings of the 42nd AES Conference on Semantic Audio*, pages 275–284, Ilmenau, Germany, 2011.

Aleksandar Stupar and Sebastian Michel. PICASSO — To Sing you must Close Your Eyes and Draw. In *Proceedings of the 34th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 715–724, Beijing, China, July 2011.

Philip Tagg and Bob Clarida. *Ten Little Title Tunes — Towards a Musicology of the Mass Media.* The Mass Media Scholar's Press, New York, USA and Montreal, Canada, 2003.

Euler C. F. Teixeira, Mauricio A. Loureiro, Marcelo M. Wanderley, and Hani C. Yehia. Motion Analysis of Clarinet Performers. *Journal of New Music Research*, July 2014.

Emiru Tsunoo, George Tzanetakis, Nobutaka Ono, and Shigeki Sagayama. Beyond Timbral Statistics: Improving Music Classification Using Percussive Patterns and Bass Lines. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4):1003–1013, May 2011.

Matthew A. Turk and Alexander P. Pentland. Face recognition using eigenfaces. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–591, Maui, Hawaii, USA, June 1991.

Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):467–476, February 2008.

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

John Unsworth. New Methods for Humanities Research. The 2005 Lyman Award Lecture, November 2005.

Julián Urbano. Information Retrieval Meta-Evaluation: Challenges and Opportunities in the Music Domain. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 609–614, Miami, USA, October 2011.

Julián Urbano, Jorge Morato, Mónica Marrero, and Diego Martín. Crowdsourcing Preference Judgments for Evaluation of Music Similarity Tasks. In *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, Geneva, Switzerland, 2010.

Zeno Vendler. *Linguistics in Philosophy*. Cornell University Press, 1967.

Raynor Vliegendhart, Martha Larson, Christoph Kofler, Carsten Eickhoff, and Johan Pouwelse. Investigating Factors Influencing Crowdsourcing Tasks with High Imaginative Load. In *Proceedings of the WSDM 2011 Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011)*, Hong Kong, China, 2011.

Avery Li-Chun Wang. An Industrial-Strength Audio Search Algorithm. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR)*, pages 7–13, Baltimore, USA, October 2003.

Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

David M. Weigl and Catherine Guastavino. User Studies in the Music Information Retrieval Literature. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 335–340, Miami, USA, 2011.

Felix Weninger, Björn Schuller, Cynthia C. S. Liem, Frank Kurth, and Alan Hanjalic. Music Information Retrieval: An Inspirational Guide to Transfer from Related Disciplines. In Meinard Müller, Masataka Goto, and Markus Schedl, editors, *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, pages 195–216. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2012.

Jason Weston, Samy Bengio, and Philippe Hamel. Multi-Tasking with Joint Semantic Spaces for Large-Scale Music Annotation and Retrieval. *Journal of New Music Research*, 40:337–348, December 2011.

Paul White. Automation For The People. *Sound on Sound*, October 2008. URL `http://www.soundonsound.com/sos/oct08/articles/leader_1008.htm`.

Brian Whitman and Ryan Rifkin. Musical Query-By-Description as a Multiclass Learning Problem. In *Proceedings of the IEEE 5th Workshop on Multimedia Signal Processing (MMSP)*, pages 153–156, St. Thomas, Virgin Islands, USA, 2002.

Frans Wiering. Meaningful Music Retrieval. In *Proceedings of the 1st Workshop on the Future of Music Information Retrieval (f(MIR)) at ISMIR 2009*, Kobe, Japan, 2009.

Frans Wiering and Anja Volk. Musicology. Tutorial slides, ISMIR, 2011.

Geraint A. Wiggins. Computer-Representation of Music in the Research Environment. In Tim Crawford and Lorna Gibson, editors, *Modern Methods for Musicology: Prospects, Proposals and Realities*, pages 7–22. Ashgate, 2009.

Geraint A. Wiggins. "I let the music speak": cross-domain application of a cognitive model of musical learning. In Patrick Rebuschat and John Williams, editors, *Statistical Learning and Language Acquisition*. Mouton De Gruyter, 2011.

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*, Discussion Forum 5:231–255, 2010.

Karthik Yadati, Martha Larson, Cynthia C. S. Liem, and Alan Hanjalic. Detecting Drops in Electronic Dance Music: Content based approaches to a socially significant music event. In *Proceedings of the 15th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 143–148, Taipei, Taiwan, 2014.

Yinsheng Zhou, Graham Percival, Xinxi Wang, Ye Wang, and Shengdong Zhao. MOGCLASS: Evaluation of a Collaborative System of Mobile Devices for Classroom Music Education of Young Children. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI)*, pages 523–532, Vancouver, Canada, 2011.

Roger Zimmermann, Elaine Chew, Sakire Arslan Ay, and Moses Pawar. Distributed Musical Performances: Architecture and Stream Management. *Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 4(2), May 2008.

ACKNOWLEDGEMENTS

What a happy coincidence it is that my thesis defense got scheduled on Thanksgiving Day. Looking back to the road towards this thesis, I truly am indebted to many people. At the risk of accidentally omitting names to mention, I feel I owe it to you all to at least try including a more comprehensive acknowledgements section at this point...

Alan, as my supervisor you will go first here. Over the past years, from a teacher you grew into a mentor to me, subsequently into a colleague, and by now I dare say into a dear friend. Thank you for the trust you put in me, which already started in my MSc years, and became particularly important at the very start of the thesis journey. Thanks for the continuous support you provided, while retaining a remarkable amount of positivity and open-mindedness, especially through the 'off-road' parts of my thesis journey, in which my funding status, organizational activities and even my geographical whereabouts were not what they normally would be for a PhD student. Whether in Delft, on the road to Luxembourg, on Barcelona hotel rooftop terraces, or over Skype while it was deep at night at at least one side of the communication line, I have always thoroughly enjoyed our extended discussions, and am looking forward to many more to come.

I would not have ended up in the current Multimedia Computing Group if as a student, I would not have been triggered by inspiring teachers and subject matter. Emile, Richard, and Inald, your courses in Image Processing, Signal Processing and Stochastic Processes in the second and third years of my BSc curriculum were the original reasons for me to gravitate towards the former ICT group. Thanks a lot for this. This makes it even more special for me to now, exactly ten years later, get to co-teach the second year BSc Signal Processing course with Inald.

Now I am a teacher myself, I would like to thank my past and present students for helping me stay sharp, and challenging me to inspire them. Alessio and Karthik, you have been teaching me at least as much as I am trying to teach you. Thank you so much for allowing me to guide you on your journey. And thanks for sharing so many great moments with me along the way, both professionally and personally.

There are many more people I am very happy to have around daily at the office in Delft. Martha, a special thanks to you, for all the interesting discussions, useful advices, and particularly for your support on my soundtrack narrative work, for which you even generously backed out of parts of your holiday. Mark, thanks for being part of my daily digital office environment, and giving so much more significance to (and practical eye-openers on) our work. Egbert, thanks for all your support to me and the group. Judith, Huijuan, Pablo, Omar, Hayley, Alessandro, and other staff colleagues, thanks for creating such an energetic work environment around me. I look very much forward to intensifying our collaborations over the coming years. Saskia, Robbert, Bart, Ruud, Caroline, and Onno, thanks for all your facilitating support over the years. Other students of the MMC group, other colleagues on the 11th floor and beyond, thanks for being there and for all the large and small activities we get to share, even if it is a cup of coffee or tea.

Yue Shi, 'Uncle' Stevan, Carsten, Christoph, Raynor, we traveled a significant part of the journey together and I hope our future roads will keep intersecting many times. Former colleagues and fellow students of the former ICT group and MSP section, thanks for having co-paved the roads towards my directions of interest. Other (former) colleagues of EEMCS and the broader university, thanks for the countless activities I got to share with you over the past years, rightfully making me feel like a TU Delft baby dinosaur by now. Special words of thanks go to former dean Daan Lenstra and current dean Rob Fastenau for their belief and support. Rob, having you as part of my audience at the Dutch Classical Talent finals was an award in itself. Speaking about awards, my gratitude also goes to the UfD Universiteitsfonds Delft for having granted me the 2012 Best PhD Candidate Award.

It was not quite trivial that I would end up as a researcher in Music Information Retrieval. For this, I should credit Rainer Typke: the press attention surrounding his thesis defense in 2007 originally made me discover the field. Huge thanks to Frans Wiering for immediately welcoming me to the symposium surrounding Rainer's defense (despite me still being a clueless BSc student at that time), and for instantly adopting me into the ISMIR community together with his colleagues and (by now former) students, in particular Bas de Haas and Peter van Kranenburg. Members (now colleagues) of ISMIR, thank you so much for providing the friendliest and most encouraging research community I could imagine. Emilia and Anja, thank you for particularly being role models to the community.

For my PhD topic, I was set on finding an own direction involving Music Information Retrieval, but in connection to the Multimedia interests in the group. The only way to realize this was to attract new, own funding. My deepest gratitude goes to Google Inc. for having saved me here. Without the Google European Doctoral Fellowship in Multimedia, the current thesis would not have existed. I also wish to thank the company for its hospitality in getting me over multiple times for internships. These were all tremendously valuable in demonstrating 'state-of-the-art' in practice and making me change my perspective on industry and Silicon Valley completely (yes, in the positive sense).

David and Beate, thanks for your university outreach work in Europe, which made the connection to Google possible for people like me. Maggie, many thanks for leading these efforts overall, and for the interesting music discussions we always got to squeeze in while I was in Mountain View. Dedicated thanks to Alfred Spector for caring about classical music organization up to the point of encouraging a dedicated internship on the topic. Doug, it was my pleasure having you as my Fellowship mentor, and being able to become a recurring insider in the team. Huge thanks to my inspiring internship supervisors Florian (in London), Rif, and Daniel, and to all the further amazing and inspiring people I met at Google Play Music, in the broader contexts of the company, and beyond. Sofia, Elizabeth, and the Tanojo family, thanks for having opened your homes to me during my California visits. Elizabeth, it was great exploring the Palo Alto area with you. Oom Hanafi, Tante Natalia, Jeremy and Naomi, thanks for simply including me as part of your family. Craig, it's been great fun visiting you at Stanford and joining you to the Stanford theatre and local concerts. Friends of the Covenant Presbyterian community, it was my pleasure to join you in making music and becoming part of the local community. I terribly miss you all, while I am happy and grateful to know that I will always have

multiple homes to (re)turn to on the West Coast.

While I was trying to unify music and multimedia research, many people guided and nudged me into the right direction. Roeland, you are fully to blame for planting the idea for the MIRUM workshop inside my head, and as such for the current increased visibility for audio, speech and music at ACM Multimedia these days. Thanks for having given this first push, and for our many music-filled interactions. Meinard, George, Doug, Steve, Bryan, Xavi, and Ye, thanks for your help in coordinating the MIRUM workshop and the subsequent ACM MM area, and for generally helping in sustaining the community. Markus, Nicola and Geoffroy, thanks for your cooperation in the MediaEval benchmark tasks. Meinard, a special extra thanks to you for involving me in your Dagstuhl seminars. The 2011 seminar has been truly important in helping me find my way and connect to colleagues. Here, I found my co-authors for Chapter 2 of this thesis, and many ad hoc music partners who even initiated me into jazz (thanks to you all!). Looking forward greatly to return to Dagstuhl in 2016.

And then there was my unplanned 'baby', the PHENICX project. It was the unforeseen obstacle preventing an earlier defense date for this thesis—yet an invaluable asset yielding knowledge, experience, new colleagues and friends, and purpose in more ways than I ever imagined. Alba, I will never forget how the two of us converted PECES into PHENICX over sleepless nights during the Easter weeks of 2012, and how you encouraged me to hang in there in the last hours until the deadline. Thank you for the abundance of proactive and positive energy from this very start, and throughout the project. While we will miss you enormously, I am happy that more fields will now have a chance to benefit from your radiance. Emilia, as mentioned before, you have been an important role model to me. I am very honored to have had a chance to co-coordinate PHENICX with you, and to share both professional and personal experiences next to this. Let's keep looking for a way to continue these exchanges in the future. Rainer, having the person who unwittingly attracted me to the Music Information Retrieval world being the initial PO for my first EU project could not have been more suitable.

Bauke, thanks a lot for your willingness to connect to parties like us (also beyond PHENICX!), and for having introduced David to me. David, thanks for having allowed us to work with your audiences, and experiment up to the Concertgebouw. Ron, your enthusiasm is exactly what we need to push PHENICX out into the world. Markus, Marko, thanks for the many interesting thought exchanges on our many shared tasks. Marcel, your support both in local and global situations has been invaluable, and I cherish the good memories surrounding the events we jointly attended and places we jointly explored afterwards. Andi, it has been great traveling the world with you, giving me an opportunity to re-explore the solo repertoire and research my own stage limits, while always benefiting from your recommendations for the after-work drinks. Everyone else at the UPF, ESMUC, JKU, OFAI, VD, and RCO, thanks enormously for having joined our mission and shared the many memorable moments.

I feel tremendously grateful for having had a chance to sustain and develop my activity as a performing musician next to the labors of this thesis, my broader academic duties, and even during my internships abroad. Even more strongly, I never would have dared to dream that my Music and Computer Science activities would become so synergetic and beneficial to one another as they are at present.

It saddens me that my former piano teachers at conservatoire, Michael Davidson and Rian de Waal, cannot witness this thesis and my current activities anymore. They are to credit for me having been able to start qualifying for a double career. Fortunately, the people who supported and guided me in my final conservatoire years are still around. Many thanks to Paul Komen and Han-Louis Meijer for having supported my path while making me a better musician with a broad artistic perspective, and to Jaring Walta for 'adopting' me at times and further broadening my musical horizon. I also am very grateful to the Dutch Classical Talent Tour & Award for having given the Magma Duo significant development possibilities, and for leading us to our current musical mentors Aleksey Igudesman and Hyung-ki Joo, who generously have been turning our musical world upside down and inside out.

Thanks to all colleagues with whom I got to share stages and rehearsals, with special thanks to my recurring musical partners: Katrien, Klarijn and Mirjam, and of course Emmy. Emmy, my dearest musical 'partner-in-crime', thank you so much for all the fun moments we got to share, and surely will share in the future. Even in the busiest of days, the Magma Duo rehearsals have always made me smile, relax and simply love what I am doing, and nothing beats the thrill and fun of jointly sharing the stage afterwards. Ger, Els, and Annelies, I have been calling Emmy my second sister, and surely you have become my second family in the meantime. Thank you very much for all your support in our endeavors, and the warmth I always find with you.

I also should mention the many interesting cooperations I had in the semi-professional music scene, most strongly in Krashna Musika and the Netherlands Student Chamber Choir. These greatly inspired me, even leading to innovations in my professional music practice and the way I perform research and dissemination (e.g., experiences around opera productions directly inspired Chapter 7 of this thesis!). I am particularly proud to see faculty lunch concerts still being held, connecting people across TU Delft. Thanks to everyone who joined me in all these associations and initiatives.

All the projects above got me many friends, but also outside of these activities, I am happy to have people to turn to. Special thanks to Fred, Peter, Henk, Lian Ien, and Jassin, for lending supportive ears to me throughout the years.

And finally, there is family. Oma Kheng en opa Sing Poen, het spijt me dat Opa Ping Lok en Oma Hiang de publicatie van dit boek niet kunnen meemaken, maar ik ben heel blij dat u er nog beiden bent. Dank u voor uw steun en liefde, en voor de risico's die u in het verleden heeft durven nemen om uiteindelijk een nieuw bestaan in Nederland op te bouwen. Had u die stap niet gezet, dan was ik hier niet geweest. Ik ben vereerd dat ik een groot deel van mijn huidige ambitie en ondernemingsdrang van u heb.

Papa en mama, sorry dat mijn activiteit vaak onnavolgbaar is geweest, en dat ik jullie vaak niet de aandacht heb kunnen geven die jullie van me zouden verwachten. Ik ben op een heel ander pad beland dan jullie je hadden voorgesteld (de verwondering geldt ook voor mijzelf), maar ik ben heel dankbaar dat jullie me hierbij toch altijd hebben gesteund. Jullie aanwezigheid en liefde betekenen meer voor me dan ik ooit in woorden zal kunnen omschrijven of in daden zal kunnen teruggeven. Bedankt voor alles.

En Vera, een beter zusje had ik me niet kunnen wensen. Als ik hier zou omschrijven wat je allemaal voor me hebt gedaan en betekend zou de dissertatie nog dikker worden dan hij al is. Dankje voor alles, en het simpele gegeven dat je er bent.

CURRICULUM VITÆ

Cynthia Cheng Sien Liem was born on February 2, 1987 in The Hague, The Netherlands. From 1998 until 2004, she followed her secondary school education at the Christelijk Gymnasium Sorghvliet, The Hague, completing her final exams with the highest final grades of the city. Perfect scores in Chemistry, French and Latin led to special awards by national associations connected to these subjects.

After this, she pursued the BSc degree (2004–2007) and MSc degree with additional honors track (2007–2009) at Delft University of Technology, which both were completed with honors. During her studies, she was the recipient of several prestigious awards: the Young Talent Incentive Prize for Computer Science of the Royal Holland Society of Sciences and Humanities 2005 (for best results in first year of study), the Lucent Global Science Scholar Award 2005, and the Google Anita Borg Scholarship 2008. During these studies, she also gained industrial experience, with research internships at Bell Labs Europe Netherlands and Philips Research, and a software engineering internship at Google UK Ltd in London.

In parallel to her computer science studies, she pursued BMus (2004–2009) and MMus (2009–2011) degrees in classical piano performance at the Royal Conservatoire in The Hague. She still is an active performer, in particular with the Magma Duo (together with Emmy Storms, violin) which has been award-winning both nationally (Laureate of the 'Dutch Classical Talent' Tour & Award 2013–2015) and internationally (First Prize and Special Prizes winner of the first international 'A Feast of Duos' duo competition in 2014 in Sion, Switzerland). Her active musicianship has been a major inspiration for her research agenda, for active efforts on making better connections between music data processing technologies and viable application scenarios and contexts, and for disseminating and (re)presenting advances in these fields to the general public.

From 2009, Cynthia Liem pursued the PhD degree at Delft University of Technology under the supervision of Prof. dr. Alan Hanjalic. She became recipient of the first Google European Doctoral Fellowship in Multimedia (2010) which supported a considerable part of her studies. Through this Fellowship, she also got connected to the team of Dr. Douglas Eck at Google Inc., leading to three software engineering internships at the Google Play Music team in Mountain View, CA, USA in 2011, 2013, and 2014. In 2012, she received the UfD—Best PhD Candidate Award at Delft University of Technology. In 2014, she made the VIVA400 shortlist of inspiring women in The Netherlands.

During her PhD studies, she pro-actively worked on increasing visibility and strength of work on Music Information Retrieval in the Multimedia research community, and pushing the paradigm of 'music as a multimedia phenomenon'. As a consequence, she initiated and co-organized the MIRUM workshop on Music Information Retrieval with User-Centered and Multimodal Strategies at ACM Multimedia (2011, 2012) which grew into a dedicated submission area at the ACM Multimedia Conference, for which she served as an area chair in 2013 and 2014. Next to this, as a MusiClef co-organizer, she

introduced the first music-related tasks in the MediaEval multimedia benchmarking initiative in 2012 and 2013.

Finally, she led acquisition efforts towards the European FP7 PHENICX project, of which she became a work package leader, dissemination coordinator, and co-coordinator of daily project management, playing an important role in the strategic agenda and public representation of this project.

Since April 2014, Cynthia Liem has entered the tenure track at Delft University of Technology as an Assistant Professor at the Multimedia Computing Group.

FULL LIST OF PUBLICATIONS

18. Mark S. Melenhorst and **Cynthia C. S. Liem**. Put the Concert Attendee in the Spotlight. A User-Centered Design and Development approach for Classical Concert Applications. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 800–806, Málaga, Spain, October 2015.

17. **Cynthia C. S. Liem** and Alan Hanjalic. Comparative Analysis of Orchestral Performance Recordings: an Image-Based Approach. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 302–308, Málaga, Spain, October 2015.

16. **Cynthia C. S. Liem**, Emilia Gómez, and Markus Schedl. PHENICX: Innovating the Classical Music Experience. In *Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME) (EU-Projects Program)*, Torino, Italy, June 2015.

15. **Cynthia C. S. Liem**. Mass Media Musical Meaning: Opportunities from the Collaborative Web. In *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 689–696, Plymouth, UK, June 2015.

14. Mark S. Melenhorst, Ron van der Sterren, Andreas Arzt, Agustín Martorell, and **Cynthia C. S. Liem**. A Tablet App to Enrich the Live and Post-Live Experience of Classical Concerts. In *Proceedings of the 3rd International Workshop on Interactive Content Consumption (WSICC) at TVX 2015*, Brussels, Belgium, June 2015.

13. Andreas Arzt, Sebastian Böck, Sebastian Flossmann, Harald Frostel, Martin Gasser, **Cynthia C. S. Liem**, and Gerhard Widmer. The Piano Music Companion. In *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI) — PAIS Conference on Prestigious Applications of Intelligent Systems*, Prague, Czech Republic, August 2014. [best demo award]

12. Emilia Gómez, Maarten Grachten, Alan Hanjalic, Jordi Janer, Sergi Jordà, Carles F. Julià, **Cynthia C. S. Liem**, Agustín Martorell, Markus Schedl, and Gerhard Widmer (alphabetical order). PHENICX: Performances as Highly Enriched aNd Interactive Concert Experiences. In *Proceedings of the SMAC Stockholm Music Acoustics Conference 2013 and SMC Sound and Music Computing Conference 2013*, Stockholm, Sweden, July 2013.

11. **Cynthia C. S. Liem**, Alessio Bazzica, and Alan Hanjalic. Looking Beyond Sound: Unsupervised Analysis of Musician Videos. In *Proceedings of the 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIA^{2}MIS)*, Paris, France, July 2013.

10. **Cynthia C. S. Liem**, Ron van der Sterren, Marcel van Tilburg, Álvaro Sarasúa, Juan J. Bosch, Jordi Janer, Mark S. Melenhorst, Emilia Gómez, and Alan Hanjalic. In-

novating the Classical Music Experience in the PHENICX Project: Use Cases and Initial User Feedback. In *Proceedings of the 1st International Workshop on Interactive Content Consumption (WSICC) at EuroITV 2013*, Como, Italy, June 2013.

9. Markus Schedl, Nicola Orio, **Cynthia C. S. Liem** and Geoffroy Peeters. A Professionally Annotated and Enriched Multimodal Data Set on Popular Music. In *Proceedings of the 4th ACM Multimedia Systems Conference (MMSys)*, pages 78–83, Oslo, Norway, February 2013.

8. **Cynthia C. S. Liem**, Martha A. Larson, and Alan Hanjalic. When Music Makes a Scene — Characterizing Music in Multimedia Contexts via User Scene Descriptions. *International Journal of Multimedia Information Retrieval*, 2(1):15–30, 2013.

7. **Cynthia C. S. Liem**, Alessio Bazzica, and Alan Hanjalic. MuseSync: Standing on the Shoulders of Hollywood. In *Proceedings of the 20th ACM International Conference on Multimedia — Multimedia Grand Challenge*, pages 1383–1384, Nara, Japan, October 2012.

6. **Cynthia C. S. Liem**, Andreas Rauber, Thomas Lidy, Richard Lewis, Christopher Raphael, Joshua D. Reiss, Tim Crawford, and Alan Hanjalic. Music Information Technology and Professional Stakeholder Audiences: Mind the Adoption Gap. In *Multimodal Music Processing*, Dagstuhl Follow-Ups vol. 3, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, pages 227–246, 2012.

5. Felix Weninger, Bjoern Schuller, **Cynthia C. S. Liem**, Frank Kurth, and Alan Hanjalic. Music Information Retrieval: An Inspirational Guide to Transfer from Related Disciplines. In *Multimodal Music Processing*, Dagstuhl Follow-Ups vol. 3, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, pages 195–216, 2012.

4. Markus Schedl, Sebastian Stober, Emilia Gómez, Nicola Orio, and **Cynthia C. S. Liem**. User-Aware Music Retrieval. In *Multimodal Music Processing*, Dagstuhl Follow-Ups vol. 3, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, pages 135–156, 2012.

3. **Cynthia C. S. Liem**, Meinard Müller, Douglas Eck, George Tzanetakis, and Alan Hanjalic. The Need for Music Information Retrieval with User-Centered and Multimodal Strategies. In *Proceedings of the 1st International ACM workshop on Music Information Retrieval with User-Centered and Multimodal Strategies (MIRUM) at ACM Multimedia*, pages 1–6, Scottsdale, USA, November 2011.

2. **Cynthia C. S. Liem** and Alan Hanjalic. Expressive Timing from Cross-Performance and Audio-based Alignment Patterns: An Extended Case Study. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 519–524, Miami, Florida, USA, October 2011.

1. **Cynthia C. S. Liem**, Alan Hanjalic, and Craig Stuart Sapp. Expressivity in Musical Timing in Relation to Musical Structure and Interpretation: A Cross-Performance, Audio-Based Approach. In *Proceedings of the 42nd International AES Conference on Semantic Audio*, pages 255–264, Ilmenau, Germany, July 2011.

ABOUT THE COVER

The cover of this thesis is based on an image from the Social Web: 'Floating Apples', published on June 4, 2007 on Flickr (https://www.flickr.com/photos/sandynelson/529250521/) by Sandy (Nelson) Maynard (https://www.flickr.com/photos/sandynelson/), and released under CC BY 2.0 (https://creativecommons.org/licenses/by/2.0/).

The image accidentally came up under search results for pictures of work by Dale Chihuly, a Tacoma-based glass artist who strongly roots his work in organic forms, shapes and colors as found in nature. I originally wanted to look for illustrations of his work as potential source material for my cover, as it makes one literally see everyday objects in new lights—a phenomenon I am advocating in this thesis with respect to the significance of user-generated videos and music performance.

This image clearly did not consider a Chihuly work. Yet, presenting many versions of glass-blown apples in an unusual context, seemingly containing sparkling jewels from afar (that triggered me; after all, I did come up with a gemstone metaphor at the start of my thesis!), it was both atmospheric and intriguing at the same time. Even more so, given the photographer's own description of the image:

> *"Floating "apples" (listed as pears) at the Tacoma Glass Museum a couple of years ago."*

I started wondering what interpretations these apple-pears were supposed to trigger. Wishing to find out more about the actual artwork depicted in the image, and why the apples were presented as pears, I ultimately discovered that the apples were part of an installation called "Blackbird in a Red Sky (a.k.a. Fall of the Blood House)" by Mildred Howard, exhibited at the Tacoma Museum of Glass from 2002–2005.

Quoting from the museum's description of Howard's work,

> *"Mildred Howard's work draws on a wide range of historical experiences and deals with family, history, home and memory. As a result, generations of story telling, which she calls collective memory, are present in her work. Howard feels that once she completes a work of art, she becomes a spectator and another story is created. Well known for her use of ordinary objects, Howard creates her installations with mixed media, often using stereotypical images, to deal with issues of domesticity, gender and race. Architectural elements, as well as found and purchased objects, are employed to create her unique visual language."*

In Howard's own description, the installation, consisting of a red house and a sea of floating apples, had deeper meaning along those lines:

> *"The humanistic referents in Blackbird in a Red Sky, also known as Fall of the Blood House—African-American history and the feminine—are important to me. I contemplate how, in an ostensibly open-ended continuum of received knowledge, personal narrative and established histories can shape understanding."*

In the framing chosen by Sandy Maynard, the house does not appear at all. Overview pictures that can be found on the Web have a very different atmosphere[5]. There also is no mentioning of pears or apples at all in these descriptions (although sources can be found speaking of "red glass apples floating in the reflecting pool"[6]).

Did the photographer and the artist interpret the work in a different way?

And does that ultimately matter?

For what we know, the photographer at least saw a situation that seemed aesthetic and interesting enough to capture. And I discovered an image that increasingly seemed to touch upon themes in my thesis and my personal road towards it, even more so as Mildred Howard turns out to be Bay Area-based.

Since the license connected to the image allowed for it, in our own turn, my sister and I decided to give some further twists to it while designing the cover. We leave it up to the curious reader to explore what these twists were, and to guess why we applied them.

[5] e.g. see `http://museumofglass.org/page.aspx?pid=478`, accessed November 4, 2015.

[6] `http://museumofglass.org/document.doc?id=21`, accessed November 4, 2015

Music is a multifaceted phenomenon. In its creation, consumption and communication, multiple modalities are at play. Next to this, music allows various ways of interpretation, is experienced as part of various everyday contexts, and is a topic under study in many different research fields, ranging from the humanities and social sciences to natural sciences, and—with the advent of the digital age—engineering as well.

In this thesis, we argue that the full potential of digital music data can only be unlocked when explicitly considering music as a multifaceted phenomenon. Adopting this view, we provide multiple novel studies and methods for problems in the Music Information Retrieval field.

A major part of the thesis is formed by the presentation of novel methods to perform data-driven analyses of multiple recorded music performances. The other major part of the thesis considers approaches for the challenge of auto-suggesting suitable soundtracks for user-generated videos. Further contributions consist of extensive positioning of the newly proposed directions in relation to existing work and known end-user stakeholder demands, leading to clear follow-up directions towards both novel research and practical impact.

ISBN 978-94-6299-238-2