Multimodal Music Processing

Edited by
Meinard Müller
Masataka Goto
Markus Schedl

DAGSTUHL
FOLLOW-UPS

Dagstuhl Follow-Ups – Vol. 3 www.dagstuhl.de/dfu

Editors

Meinard Müller
Saarland University
and Max-Planck Institut
für Informatik
meinard@mpi-inf.mpg.de

Masataka Goto
National Institute of
Advanced Industrial
Science and Technology (AIST)
m.goto@aist.go.jp

Markus Schedl
Department of
Computational Perception
Johannes Kepler University
markus.schedl@jku.at

ACM Classification 1998
H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems

ISBN 978-3-939897-37-8

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at http://www.dagstuhl.de/dagpub/978-3-939897-37-8.

Publication date
April, 2012

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Cover graphic
The painting of Ludwig van Beethoven was drawn by Joseph Karl Stieler (1781–1858). The photographic reproduction is in the public domain.

Digital Object Identifier: 10.4230/DFU.Vol3.11041.i

ISBN 978-3-939897-37-8 **ISSN 1868-8977** **http://www.dagstuhl.de/dfu**

DFU – Dagstuhl Follow-Ups

The series *Dagstuhl Follow-Ups* is a publication format which offers a frame for the publication of peer-reviewed papers based on Dagstuhl Seminars. DFU volumes are published according to the principle of Open Access, i.e., they are available online and free of charge.

ISSN 1868-8977

www.dagstuhl.de/dfu

Contents

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl
Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

Preface

Music can be described, represented, and experienced in various ways and forms. For example, music can be described in textual form not only supplying information on composers, musicians, specific performances, or song lyrics, but also offering detailed descriptions of structural, harmonic, melodic, and rhythmic aspects. Music annotations, social tags, and statistical information on user behavior and music consumption are also obtained from and distributed on the world wide web. Furthermore, music notation can be encoded in text-based formats such as MusicXML, or symbolic formats such as MIDI. Beside textual data, increasingly more types of music-related multimedia data such as audio, image or video data are widely available. Because of the proliferation of portable music players and novel ways of music access supported by streaming services, many listeners enjoy ubiquitous access to huge music collections containing audio recordings, digitized images of scanned sheet music and album covers, and an increasing number of video clips of music performances and dances.

This volume is devoted to the topic of multimodal music processing, where both the availability of multiple, complementary sources of music-related information and the role of the human user is considered. Our goals in producing this volume are two-fold: Firstly, we want to spur progress in the development of techniques and tools for organizing, analyzing, retrieving, navigating, recommending, and presenting music-related data. To illustrate the potential and functioning of these techniques, many concrete application scenarios as well as user interfaces are described. Also various intricacies and challenges one has to face when processing music are discussed. Our second goal is to introduce the vibrant and exciting field of music processing to a wider readership within and outside academia. To this end, we have assembled thirteen overview-like contributions that describe the state-of-the-art of various music processing tasks, give numerous pointers to the literature, discuss different application scenarios, and indicate future research directions. Focusing on general concepts and supplying many illustrative examples, our hope is to offer some valuable insights into the multidisciplinary world of music processing in an informative and non-technical way.

When dealing with various types of multimodal music material, one key issue concerns the development of methods for identifying and establishing semantic relationships across various music representations and formats. In the first contribution, Thomas *et al.* discuss the problem of automatically synchronizing two important types of music representations: sheet music and audio files. While sheet music describes a piece of music visually using abstract symbols (e. g., notes), audio files allow for reproducing a specific acoustic realization of a piece of music. The availability of such linking structures forms the basis for novel interfaces that allow users to conveniently navigate within audio collections by means of the explicit information specified by a musical score. The second contribution on lyrics-to-audio alignment by Fujihara and Goto deals with a conceptually similar task, where the objective is to estimate a temporal relationship between lyrics and an audio recording of a given song. Locating the lyrics (text-based representation) within a singing voice (acoustic representation) constitutes a challenging problem requiring methods from speech as well as music processing. Again, to highlight the importance of this task, various Karaoke and retrieval applications are described.

The abundance of multiple information sources does not only open up new ways for music navigation and retrieval, but can also be used for supporting and improving the analysis of music data by exploiting cross-modal correlations. The next three contributions discuss such

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl
Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

multimodal approaches for music analysis. Essid and Richard first give an overview of general fusion principles and then discuss various case studies that highlight how video, acoustic, and sensor information can be fused in an integrated analysis framework. For example, the authors show how visual cues can be used to support audio-based drum transcription. Furthermore, in the case study of dance scene analysis various types of motion representations (e. g. obtained from inertial sensors or depth image sensors) are combined with video and audio representations. Konz and Müller show in their contribution how the harmonic analysis of audio recording can be improved and stabilized by exploiting multiple versions of the same piece of music. Using a late-fusion approach by analyzing the harmonic properties of several audio versions synchronously, the authors show that consistencies across several versions indicate harmonically stable passages in the piece of music, which may have some deeper musical meaning. Finally, Ewert and Müller show how additional note information as specified by a musical score can be exploited to support the task of source separation. Since such sources, which may correspond to a melody, a bassline, a drum track, or an instrument track, are mixed into monaural or stereo audio signals and highly correlated in the musical context, the problem generally becomes intractable. Here, the additional score information can be employed to alleviate and guide the separation process.

In the next two contributions, the potential of the multimodal analysis techniques are highlighted by means of different interactive application scenarios. Dittmar *et al.* show how techniques such as music transcription and sound separation open up new possibilities for various music learning, practicing, and gaming applications. In particular, a music software is presented which provides the entertainment and engagement of music video games while offering appropriate methods to develop musical skills. This software also offers functionalities that allow users to create personalized content for the game, e. g., by generating solo and accompaniment track from user-specified audio material. Dannenberg addressed in his contribution the problem of creating computer music systems that can perform live music in association with human performers. Besides the above mentioned synchronization and linking techniques, this scenario requires advanced real-time music analysis and synthesis techniques that allow the system to react to a human performance in an intelligent way.

Besides the music processing techniques and their applications as discussed so far, the problem of finding and retrieving relevant information from heterogenous and distributed music collections has substantially gained importance during the last decade. As exposed in the subsequent three contributions, the term "multimodality" can be recognized at several levels in the retrieval context. For example, one may consider different types of textual, acoustic, or visual representations of music. Or one may also consider different modalities to access music collections – query-by-example, direct querying, browsing, metadata-based search, visual user interfaces, just to name a few. The contribution by Schedl *et al.* gives an overview of various aspects of multimodal music retrieval with a particular focus on the issue on how to build personalized systems that particularly address the user's interest and behavior. In particular various relations between computational features and the human music perception are discussed, accounting for user-centered aspects such as similarity, diversity, familiarity, hotness, recentness, novelty, serendipity, and transparency. The contribution by Grosche *et al.* approaches the topic of music information retrieval from another perspective. In the case that textual descriptions are not available one requires retrieval strategies which only access the contents of the raw audio material. The authors give an overview of various content-based retrieval approaches that follow the query-by-example paradigm. Based on the principles of granularity and specificity, various notions and levels of similarity used to compare different audio recordings (or fragments) are discussed. Müller and Driedger

illustrate how various content-based analysis and retrieval techniques come into play and act together when considering a data-driven application scenario for generating sound tracks. Here, the objective is to create computer-assisted tools that allow users to easily and intuitively generate aesthetically appealing music tracks for a given multimedia stream such as a computer game or slide show.

The last three contributions of this volume reflect on the kind of role music processing has played in the past and offer a few thoughts on challenges, open problems, and future directions. As noted by Weninger *et al.*, the relatively young fields of music processing and music information retrieval have been influenced by neighboring domains in signal processing and machine learning, including automatic speech recognition, image processing and text information retrieval. In their contribution, the authors give various examples for methodology transfer, show parallel developments in the different domains, and indicate how neighboring fields may now benefit from the music domain. In a stimulating and provocative contribution, Goto describes his visions on how computed-based music processing methods may help to generate new music, to predict music trends, and to enrich our daily lives. Picking up some recent developments in Japan, various grand challenges are presented that not only indicate future research directions but also should help to increase both the attraction and social impact of research in multimodal music processing and music information retrieval. In the final contribution, Liem *et al.* reflect on the kind of impact that music processing has had across disciplinary boundaries and discuss various technology adoption issues that were experienced with professional music stakeholders in audio mixing, performance, musicology and sales industry. The music domain offers many possibilities for truly cross-disciplinary collaboration and technology. However, in order to achieve this, careful consideration of the users' actual need as well as an investment in understanding the involved communities will be essential.

This volume, which is based on our Dagstuhl seminar on "Multimodal Music Processing" held in January 2011, is the result of the work by many people. First of all, we thank the authors for their contributions as well as the reviewers for their valuable feedback. We are grateful to the *Cluster of Excellence on Multimodal Computing and Interaction* (MMCI) at Saarland University for their support. We highly appreciate and wish to thank the Dagstuhl board and the Dagstuhl office for supporting us in having the seminar. In particular, we want to thank Marc Herbstritt, who was extremely helpful with his advice and active support in preparing and editing this volume. Thank you very much.

March 2012 *Meinard Müller, Masataka Goto, and Markus Schedl*

List of Authors

Jakob Abeßer
Semantic Music Technologies Group,
Fraunhofer IDMT
Ilmenau, Germany
abr@idmt.fraunhofer.de

Estefanía Cano
Semantic Music Technologies Group,
Fraunhofer IDMT
Ilmenau, Germany
cano@idmt.fraunhofer.de

Michael Clausen
Department of Computer Science III,
University of Bonn
Bonn, Germany
clausen@cs.uni-bonn.de

Tim Crawford
Department of Computing, Goldsmiths,
University of London
London, United Kingdom
t.crawford@gold.ac.uk

Roger B. Dannenberg
Carnegie Mellon University
Pittsburgh, USA
rbd@cs.cmu.edu

Christian Dittmar
Semantic Music Technologies Group,
Fraunhofer IDMT
Ilmenau, Germany
dmr@idmt.fraunhofer.de

Jonathan Driedger
Saarland University and Max-Planck Institut
für Informatik
Saarbrücken, Germany
driedger@mpi-inf.mpg.de

Slim Essid
Institut Télécom, Télécom ParisTech,
CNRS-LTCI
Paris, France
Slim.Essid@telecom-paristech.fr

Sebastian Ewert
Department of Computer Science III,
University of Bonn
Bonn, Germany
ewerts@iai.uni-bonn.de

Christian Fremerey
Department of Computer Science III,
University of Bonn
Bonn, Germany
fremerey@cs.uni-bonn.de

Hiromasa Fujihara
National Institute of Advanced Industrial
Science and Technology (AIST)
Tsukuba, Japan
h.fujihara@aist.go.jp

Masataka Goto
National Institute of Advanced Industrial
Science and Technology (AIST)
Tsukuba, Japan
m.goto@aist.go.jp

Emilia Gómez
Music Technology Group, Universitat
Pompeu Fabra
Barcelona, Spain
emilia.gomez@upf.edu

Fabien Gouyon
Institute for Systems and Computer
Engineering, University of Porto
Porto, Portugal
fgouyon@inescporto.pt

Sascha Grollmisch
Semantic Music Technologies Group,
Fraunhofer IDMT
Ilmenau, Germany
goh@idmt.fraunhofer.de

Peter Grosche
Saarland University and Max-Planck Institut
für Informatik
Saarbrücken, Germany
pgrosche@mpi-inf.mpg.de

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl
Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

Alan Hanjalic
Multimedia Information Retrieval Lab, Delft University of Technology
Delft, The Netherlands
a.hanjalic@tudelft.nl

Verena Konz
Saarland University and Max-Planck Institut für Informatik
Saarbrücken, Germany
vkonz@mpi-inf.mpg.de

Frank Kurth
Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie FKIE
Wachtberg, Germany
frank.kurth@fkie.fraunhofer.de

Richard Lewis
Department of Computing, Goldsmiths, University of London
London, United Kingdom
richard.lewis@gold.ac.uk,

Thomas Lidy
Information Management and Preservation Lab, Vienna University of Technology
Vienna, Austria
lidy@ifs.tuwien.ac.at

Cynthia C. S. Liem
Multimedia Information Retrieval Lab, Delft University of Technology
Delft, The Netherlands
c.c.s.liem@tudelft.nl

Meinard Müller
Saarland University and Max-Planck Institut für Informatik
Saarbrücken, Germany
meinard@mpi-inf.mpg.de

Nicola Orio
Department of Information Engineering, University of Padova
Padova, Italy
orio@dei.unipd.it

Christopher Raphael
School of Informatics, Indiana University
Bloomington, USA
craphael@indiana.edu

Andreas Rauber
Information Management and Preservation Lab, Vienna University of Technology
Vienna, Austria
rauber@ifs.tuwien.ac.at

Joshua D. Reiss
Centre for Digital Music, Queen Mary, University of London
London, United Kingdom
josh.reiss@eecs.qmul.ac.uk

Gaël Richard
Institut Télécom, Télécom ParisTech, CNRS-LTCI
Paris, France
Gael.Richard@telecom-paristech.fr

Joan Serrà
Artificial Intelligence Research Institute (IIIA-CSIC)
Barcelona, Spain
jserra@iiia.csic.es

Markus Schedl
Department of Computational Perception, Johannes Kepler University
Linz, Austria
markus.schedl@jku.at

Björn Schuller
Technische Universität München
München, Germany
schuller@tum.de

Sebastian Stober
Data & Knowledge Engineering Group, Otto-von-Guericke-Universität
Magdeburg, Germany
stober@ovgu.de

Verena Thomas
Department of Computer Science III, University of Bonn
Bonn, Germany
thomas@cs.uni-bonn.de

Felix Weninger
Technische Universität München
München, Germany
weninger@tum.de

Linking Sheet Music and Audio – Challenges and New Approaches

Verena Thomas[1], Christian Fremerey[*2], Meinard Müller[†3], and Michael Clausen[4]

1,2,4 University of Bonn, Department of Computer Science III
Römerstr. 164, 53117 Bonn, Germany
{thomas,fremerey,clausen}@cs.uni-bonn.de

3 Saarland University and MPI Informatik
Campus E1-4, 66123 Saarbrücken, Germany
meinard@mpi-inf.mpg.de

Abstract

Score and audio files are the two most important ways to represent, convey, record, store, and experience music. While score describes a piece of music on an abstract level using symbols such as notes, keys, and measures, audio files allow for reproducing a specific acoustic realization of the piece. Each of these representations reflects different facets of music yielding insights into aspects ranging from structural elements (e. g., motives, themes, musical form) to specific performance aspects (e. g., artistic shaping, sound). Therefore, the simultaneous access to score and audio representations is of great importance. In this paper, we address the problem of automatically generating musically relevant linking structures between the various data sources that are available for a given piece of music. In particular, we discuss the task of sheet music-audio synchronization[1] with the aim to link regions in images of scanned scores to musically corresponding sections in an audio recording of the same piece. Such linking structures form the basis for novel interfaces that allow users to access and explore multimodal sources of music within a single framework. As our main contributions, we give an overview of the state-of-the-art for this kind of synchronization task, we present some novel approaches, and indicate future research directions. In particular, we address problems that arise in the presence of structural differences and discuss challenges when applying optical music recognition to complex orchestral scores. Finally, potential applications of the synchronization results are presented.

1998 ACM Subject Classification H.5.1 Multimedia Information Systems, H.5.5 Sound and Music Computing, I.5 Pattern Recognition, J.5 Arts and Humanities–Music

Keywords and phrases Music signals, audio, sheet music, music synchronization, alignment, optical music recognition, user interfaces, multimodality

Digital Object Identifier 10.4230/DFU.Vol3.11041.1

1 Introduction

Significant advances in data storage, data acquisition, computing power, and the worldwide web are among the fundamental achievements of modern information technology. This

* Christian Fremerey is now with Steinberg Media Technologies GmbH, Germany.
† Meinard Müller has been supported by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI). He is now with Bonn University, Department of Computer Science III, Germany.
[1] We use the term *sheet music* as equivalent to scanned images of music notation while *score* refers to music notation itself or symbolic representations thereof.

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 1–22

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

technological progress opened up new ways towards solving problems that appeared nearly unsolvable fifty years ago. One such problem is the long-term preservation of our cultural heritage. Libraries, archives, and museums throughout the world have collected vast amounts of precious cultural material. The physical objects are not only difficult to access, but also threatened from decay. Therefore, numerous national and international digitization initiatives have been launched with the goal to create digital surrogates and to preserve our cultural heritage.[2] However, generating and collecting digitized surrogates represents only the beginning of an entire process chain that is needed to avoid digital graveyards. To make the digitized data accessible, one requires automated methods for processing, organizing, annotating, and linking the data. Furthermore, intuitive and flexible interfaces are needed that support a user in searching, browsing, navigating, and extracting useful information from a digital collection.

In this paper, we address this problem from the perspective of a digital music library project,[3] which has digitized and collected large amounts of Western classical music. Such collections typically contain different kinds of music-related documents of various formats including text, symbolic, audio, image, and video data. Three prominent examples of such data types are sheet music, symbolic score data, and audio recordings. Music data is often digitized in some systematic fashion using (semi-)automatic methods. For example, entire sheet music books can be digitized in a bulk process by using scanners with automatic page turners. This typically results in huge amounts of high-resolution digital images stored in formats such as TIFF or PDF. One can further process the image data to obtain symbolic music representations that can be exported into formats such as MusicXML,[4] LilyPond,[5] or MIDI.[6] This is done by using *optical music recognition* (OMR), the musical equivalent to optical character recognition (OCR) as used in text processing. Symbolic music representations and MIDI files are also obtained from music notation software or from electronic instruments. Last but not least, modern digital music libraries contain more and more digitized audio material in form of WAV or MP3 files. Such files are obtained by systematically ripping available CD collections, converting tape recordings, or digitizing old vinyl recordings.

As a result of such systematic digitization efforts, one often obtains data sets that contain items of a single type,[7] see also Figure 1. For example, scanning entire sheet music books results in a collection of image files, where each file corresponds to a specific page. Or, ripping a data set of CD recordings, one obtains a collection of audio files, where each file corresponds to an audio track. In the case of digitizing a vinyl recording, a track covers an entire side of the recording that may comprise several pieces of music.[8] In order to make the data accessible in a user-friendly and consistent way, various postprocessing steps are required. For example, the scanned pages of sheet music need to be pooled, cut, or combined to form musically meaningful units such as movements or songs. Furthermore, these units

[2] For example, the project *Presto Space* (`http://www.prestospace.org`) or the internet portal *Europeana* (`http://www.europeana.eu`).
[3] PROBADO, for more information we refer to `http://www.probado.de/en_home.html`.
[4] `http://www.recordare.com/musicxml`
[5] `http://lilypond.org`
[6] `http://www.midi.org`
[7] For example, the *Archival Sound Recordings* of the British Library (`http://sounds.bl.uk`), the Chopin Early Editions (`http://chopin.lib.uchicago.edu`), or the *Munich Digitization Center* of the Bavarian State Library (`http://bsb-mdz12-spiegel.bsb.lrz.de/~mdz`).
[8] The notion of a piece of music usually refers to individual movements or songs within bigger compositions. However, the particular segmentation applied by music libraries can vary.

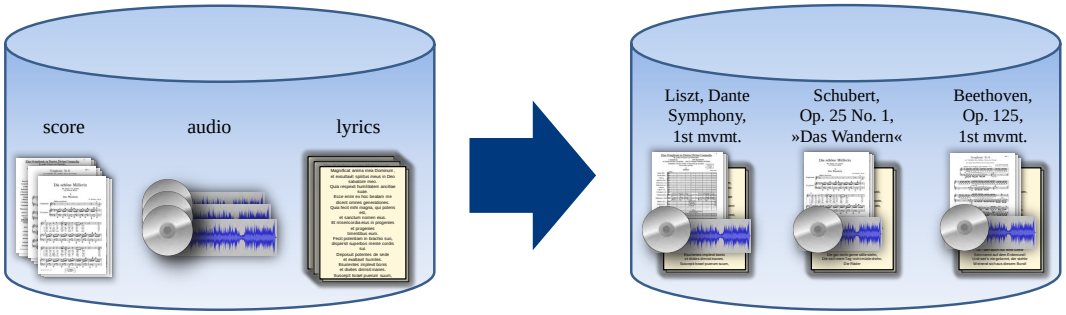
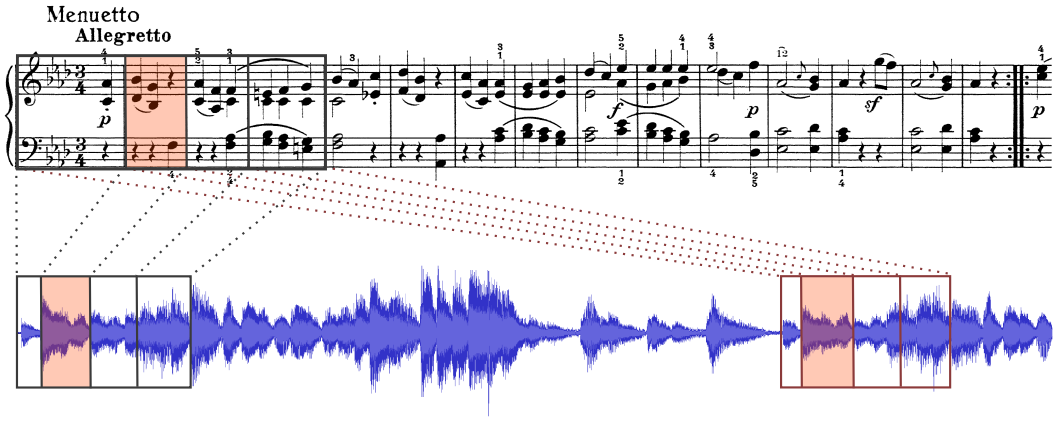


Figure 1 Change from a document and document type centered data collection (**left**) to an arrangement focusing on pieces of music (**right**).

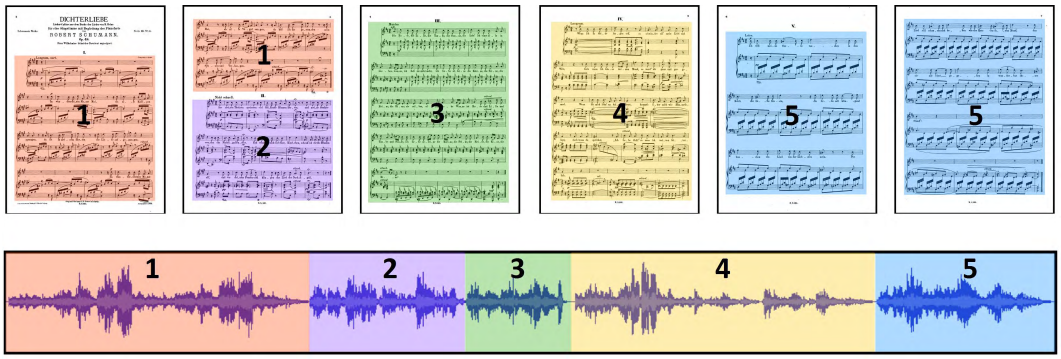
need to be assigned to the respective musical work and annotated accordingly. Similarly, audio tracks are to be identified and trimmed to meet certain standards and conventions. Finally, suitable metadata needs to be attached to the digitized documents. When trying to automate the stated postprocessing steps for real-world music collections, they become challenging research problems. The main issues are the inconsistency and the complexity of the given data. For instance, sheet music contains a lot of textual metadata but its extraction and proper interpretation are non-trivial tasks (e.g., *Allegro* can likewise constitute a tempo instruction or the name of a piece of music, see [25] for further details).

The availability of accurate metadata is essential for organizing and indexing huge music collections. For example, searching for the keywords “Beethoven” and “Op. 125”, one should be able to retrieve all documents that refer to Beethoven Symphony No. 9. In this way, suitable metadata information allows for re-arranging the music documents to obtain a data collection, where all versions that refer to the same piece of music are compiled irrespective of their format or modality, see Figure 1. However, such a document-level compilation of musically related versions constitutes only the first step towards a comprehensive system for multimodal music navigation and browsing. In the next step, one requires linking structures that reveal the musical relations within and across the various documents at a lower hierarchical level. For example, such a linking structure may reveal the musical correspondence between notes depicted in a scanned sheet music document and time positions in an audio recording of the same piece of music. Such links would then allow for a synchronous display of the audible measure in the sheet music representation during the playback of a music recording. Similarly, in a retrieval scenario, a musical theme or passage could be marked in the image domain to retrieve all available music recordings where this theme or passage is played.

In this paper, we address the problem of how suitable linking structures between different versions of the same piece of music can be computed in a fully automated fashion. In particular, we focus on the multimodal scenario of linking sheet music representations with corresponding audio representations, a task we also refer to as *sheet music-audio synchronization*, see Figure 2a. In Section 2, we give an overview of an automated synchronization procedure and discuss various challenges that arise in the processing pipeline. One major step in this pipeline is to extract explicit note events from the digitized sheet music images by using OMR. Even though there is already various commercial OMR software on the market for many years, the robust extraction of symbolic information is still problematic for complex scores. Some of these challenges are discussed in Section 3. In particular, certain extraction errors have severe consequences, which may lead to erroneous assignments of entire instrument tracks or



(a) Score-audio synchronization on the measure-level. Time segments in the audio stream are mapped to individual measures in the score representation. The depicted audio track contains a repetition. Therefore, the according score measures have to be mapped to both audio segments.



(b) Score-audio mapping on the detail level of pieces of music. The score and the audio data are segmented into individual pieces of music. Afterwards, the correct score-audio pairs have to be determined.

Figure 2 Examples for score-audio synchronization on different detail levels.

to deviations in the global music structure. In Section 4, we discuss common computational approaches to sheet music-audio synchronization and present various strategies how the resulting global differences between documents can be handled within the synchronization pipeline. Finally, in Section 5, we describe some applications and novel interfaces that are based on synchronization results. We conclude the paper with an outlook on future work. A discussion of relevant work can be found in the respective sections.

2 Task Specification

The goal of music synchronization is the generation of semantically meaningful bidirectional mappings between two music documents representing the same piece of music. Those documents can be of the same data type (e.g., audio-audio synchronization) or of different data types (e.g., score-audio synchronization or lyrics-audio synchronization). In the case of score-audio synchronization the created linking structures map regions in a musical score, e.g., pages or measures, to semantically corresponding sections in an audio stream (see Figure 2).

Although the task of score-audio synchronization appears to be straightforward, there exist several aspects along which the task and its realization can vary (see Figure 3). The

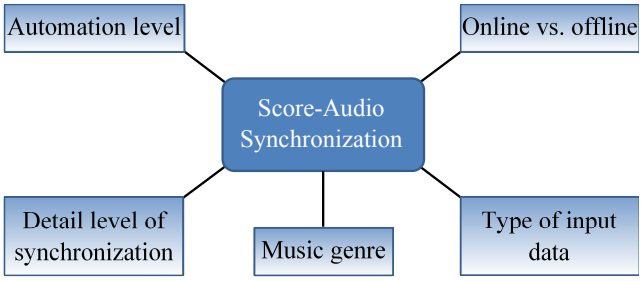


Figure 3 Aspects of score-audio synchronization.

particular choice of settings with respect to these aspects is always influenced by the intended application of the synchronization results.

The first choice concerns the sought detail level or granularity of the synchronization. A very coarse synchronization level would be a mapping between score and audio sections representing the same piece of music, see Figure 2b (e.g., *Neue Mozart-Ausgabe*[9]). This type of alignment is also referred to as *score-audio mapping*. Finer detail levels include page-wise [2, 21], system-wise, measure-wise [34], or note-wise [8, 46] linking structures between two music documents. The choice of granularity can in turn affect the level of automation. The manual annotation of the linking structure might be achievable for page-wise synchronizations. However, for finer granularities semi-automated or automated synchronization algorithms would be preferable. While automatic approaches do not need (and also not allow) any user interaction, in semi-automatic approaches some user interaction is required. However, the extent of the manual interaction can vary between manually correcting a proposed alignment on the selected detail level and correcting high-level aspects (e.g., the repeat structure) before recalculating the alignment. The selected automation level obviously also depends on the amount of data to be processed. For a single piece of music given only one score and one audio interpretation, a full-fledged synchronization algorithm might not be required. But, for the digitized music collection of a library, manual alignment becomes impossible. Finally, reliability or accuracy requirements also take part in the automation decision.

Another huge differentiation concerns the runtime scenario. In *online* synchronization, the audio stream is only given up to the current playback position and the synchronization should produce an estimation of the current score position in real-time. There exist two important applications of online score-audio synchronization techniques, namely *score following* and *automated accompaniment* [13, 17, 36, 37, 46, 48, 54]. The real-time requirements of this task turn local deviations between the score and the audio into a hard problem. Furthermore, recovery from local synchronization errors is problematic. In contrast, in *offline* synchronization the complete audio recording and the complete score data are accessible throughout the entire synchronization process [34, 42]. Also, the computation is not required to run in real-time. Due to the loosened calculation time requirements and the availability of the entire audio and score data during calculation, offline synchronization algorithms usually achieve higher accuracies and are more robust with regard to local deviations in the input data. The calculated linking structures can afterwards be accessed to allow for, e.g., score-based navigation in audio files.

The genre/style of the music to be synchronized also influences the task of score-audio synchronization. While Western classical music and most popular music feature strong

[9] `http://www.nma.at`

melodic/harmonic components other music styles, like African music, may mainly feature rhythmic drumming sounds. Obviously, using harmonic information for the synchronization of rhythmic music will prove ineffective and therefore different approaches have to be employed.

The type of input data—more precisely the score representation—constitutes the last aspect of score-audio synchronization. The score data can either be available as scanned images of music notation (i.e., sheet music) or as symbolic score (e.g., MIDI or MusicXML). Obviously, the choice of score input affects the type of challenges to be mastered during synchronization. While symbolic score representations are usually of reasonable quality and the extraction of the individual music events is straightforward, some sort of rendering is required to present the score data. In contrast, sheet music already provides a visualization. But the music information needs to be reconstructed from the image data before the linking structures can be calculated. OMR systems approach this task and achieve high reconstruction rates for printed Western music. Nevertheless, the inclusion of OMR into the synchronization process may result in defective symbolic score data (see Section 3). Usually, the errors are of mainly local nature. Thus, by choosing a slightly coarser detail level (e.g., measure level) sound synchronization results can be achieved. For a differentiation between these two types of input data, the term *sheet music-audio synchronization* is often utilized if scanned images are given as score input.

Various researchers are active in the field of score-audio synchronization and work on all settings of the listed aspects has been reported. Considering all aspects and their specific challenges would go beyond the scope of this paper. Instead, we focus on the task of automated offline sheet music-audio synchronization for Western classical music producing linking structures on the measure level. Furthermore, the processing of large music collections should be possible.

The basic idea in most score-audio synchronization scenarios is to transform both input data types into a common mid-level representation. These data streams can then be synchronized by applying standard alignment techniques, see Section 4 for an overview. Independent of the selected approach, one has to cope with the following problems to get reasonable synchronization results:

- **Differences in structure:** A score can contain a variety of symbols representing jump instructions (e.g., repeat marks, segno signs, or keywords such as *da capo*, *Coda*, or *Fine*, see Figure 4). While OMR systems are capable of detecting repeat marks, they often fail to reliably detect most other textual jump instructions in the score. Therefore, the correct playback sequence of the measures cannot be reconstructed. However, even if all jump instructions are correctly recognized, the audio recording may reveal additional repeats or omissions of entire passages notated in the score. Again, the given sequence of measures does not coincide with the one actually played in the audio recording. Such structural differences lead to major challenges in score-audio synchronization.

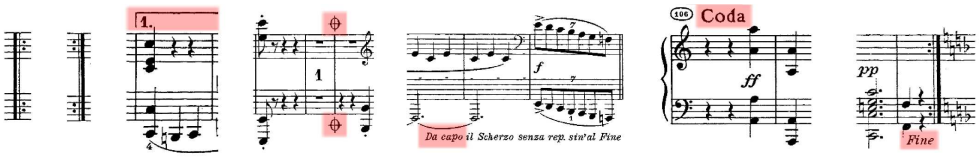


Figure 4 Examples of jump indicators used in music notation (adapted from [25]).

- **Differences between music representations:** Score pages and audio recordings represent a piece of music on different levels of abstraction and capture different facets of the music. One example is the tempo. Music notation may provide some written

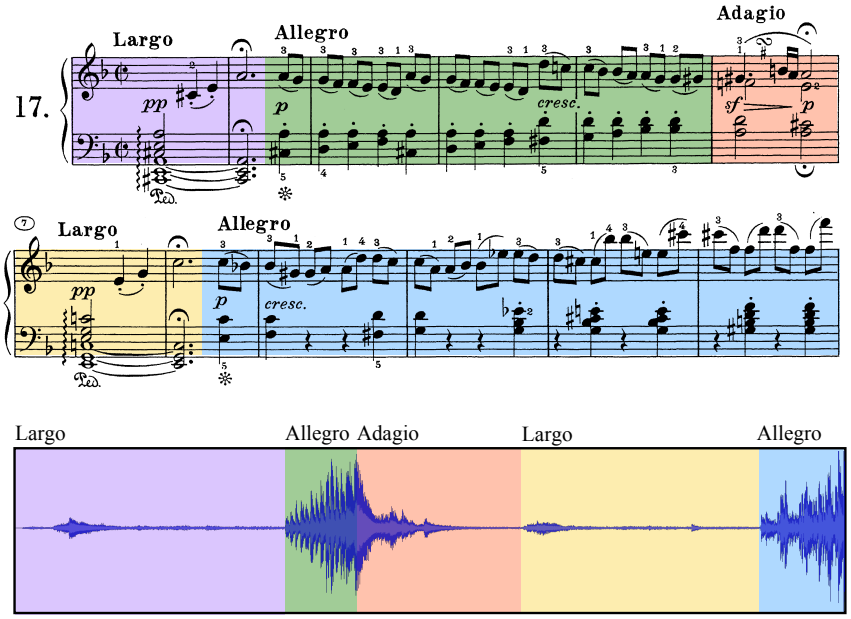


Figure 5 Extract of Beethoven's *Piano Sonata No. 17* (publisher: *Henle Verlag*, pianist: V. Ashkenazy). In the first nine measures alone four substantial tempo changes are performed. Thus, the duration of the measures in the audio recording varies significantly. However, in the score only vague instructions are available that result at best in an approximation of the intended tempo changes.

information on the intended tempo of a piece of music and tempo changes therein (e.g., instructions such as *Allegro* or *Ritardando*). However, those instructions provide only a rough specification of the tempo and leave a lot of space for interpretation. Therefore, different performers might deviate significantly in their specific tempo choices. In addition, most musicians even add tempo changes that are not specified by the score to emphasize certain musical passages. For an example we refer to Figure 5.

The differences in the loudness of instruments and the loudness variations during the progression of a piece of music are further important characteristics of a given performance. Just like tempo, loudness is notated only in a very vague way and OMR systems often fail to detect the few available instructions. Similarly, music notation only provides timbre information through instrument labels. Therefore, timbre-related sound properties such as instrument-dependent overtone energy distributions are not explicitly captured by the score.

In conclusion, in view of practicability, score-audio synchronization techniques need to be robust towards variations in tempo, loudness, and timbre to deal with the mentioned document type related differences.

- **Errors in the input data:** As already mentioned, OMR is not capable of reconstructing the score information perfectly. The errors introduced by OMR can be divided into local and global ones. Local errors concern, e.g., misidentifications of accidentals, missed notes, or wrong note durations. In contrast, examples for global errors are errors in the detection of the musical key or the ignorance of transposing instruments. Further details will be presented in Section 3. While for sheet music, errors are introduced during the reconstruction from the scanned images, the audio recordings themselves can be erroneous. The performer(s) may locally play some wrong notes or a global detuning

occurred. For Western classical music a tuning of 440 Hz for the note $A4$ was defined as standard. However, most orchestras slightly deviate from this tuning.[10] Furthermore, for Baroque music a deviation by a whole semitone is common.

- **Sheet music-audio mapping:** Especially in library scenarios, the goal is not the synchronization of one piece of music. Usually, the input consists of whole sheet music books and whole CD collections. Therefore, before calculating the linking structures, the score and the audio data need to be segmented into individual pieces of music. As the order in the sheet music books and on the CDs might differ, a mapping on this granularity level needs to be created before the actual synchronizations can be calculated.

Although we focus on sheet music-audio synchronization in this contribution, most of the mentioned problems also exist for other score-audio synchronization variants.

3 Optical Music Recognition

Similarly to optical character recognition (OCR) with the goal to reconstruct the textual information given on scanned text pages, optical music recognition (OMR) aims at restoring musical information from scanned images of sheet music. But, the automatic reconstruction of music notation from scanned images has to be considered much harder than OCR. Music notation is two-dimensional, contains more symbols, and those symbols mostly overlap with the staves. A large number of approaches to OMR has been proposed and several commercial and non-commercial OMR systems are available today. Three more popular commercial systems are SharpEye,[11] SmartScore,[12] and PhotoScore.[13] All of them operate on common Western classical music. While the former two only work for printed sheet music, PhotoScore also offers the recognition of handwritten scores. Two prominent examples for non-commercial OMR systems are Gamera[14] and Audiveris.[15] While Audiveris is not competitive in terms of recognition rates, Gamera is actually a more general tool for image analysis. Therefore, Gamera requires training on the data to be recognized to yield adequate recognition results. Since the introduction of OMR in the late 1960s [45] many researchers worked in the field and relevant work on the improvement of the recognition techniques has been reported. For further information, we refer to the comprehensive OMR bibliography by Fujinaga [29].

As with score-audio synchronization, there are three factors that affect the difficulty of the OMR task and the selection of the pursued approach. First, there exist different types of scores (e.g., medieval notation, modern notation or lute tablatures) that differ significantly in their symbol selection and their basic layout. Therefore, the type of music notation present on the images has to be considered. Second, the transcription format is of influence. Printed score is regular and usually well formatted while handwritten score can be rather unsteady and scrawly. Additionally, crossing outs, corrections, and marginal notes make the interpretation of handwritten scores even more challenging. Finally, the envisioned application of the resulting symbolic representation influences the required precision. OMR results intended for playback or score rendering have to present a much higher accuracy on the note level than a reconstruction serving as score representation during sheet music-audio synchronization on the measure level (see Section 4). In the first scenario, most OMR systems

[10] List of standard pitches in international orchestras: `http://members.aon.at/fnistl/`
[11] `http://www.music-scanning.com`
[12] `http://www.musitek.com`
[13] `http://www.sibelius.at/photoscore.htm`
[14] `http://gamera.informatik.hsnr.de`
[15] `http://audiveris.kenai.com`

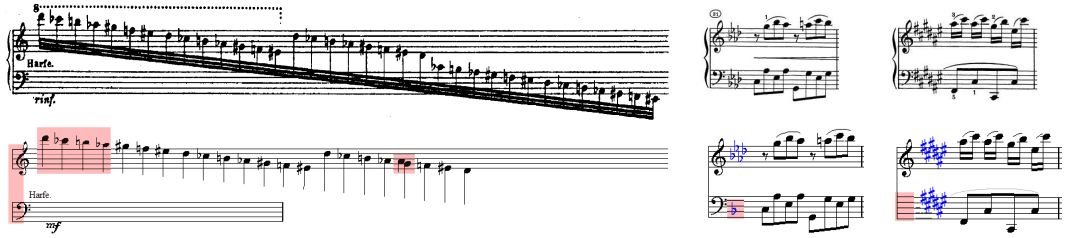


Figure 6 Examples of common OMR errors. **Left:** Besides wrong note durations and an accidental mistaken for a note, the staff system was split into two systems. **Middle:** The key signature was not correctly recognized for the lower staff. **Right:** In the lower staff, the clef was not detected.

support the creation of an initial approximation of a symbolic representation and provide user interfaces for manual correction.

Several studies on the performance of OMR systems and the types of errors that occur were conducted [10, 11, 12, 25]. Those studies show that OMR systems vary with regard to their strengths and weaknesses. However, the types or classes of recognition errors are the same for all systems. Some examples of common errors are given in Figure 6. Most of those errors are of a local nature and concern individual music symbols or small groups thereof. Examples are articulation marks, ornaments, accidentals, dynamics, and note durations that are mistaken for some other symbol or missed altogether. In the context of sheet music-audio synchronization those local errors are less severe because the applied synchronization methods are capable of managing local deviations between the two sequences to be aligned. In contrast, several types of recognition errors, influencing larger areas of the score, exist. Those might include wrong time signatures, missed clefs, wrong key signatures, staff systems being split up (e.g., due to arpeggios traveling through several staves or due to textual annotations disrupting the vertical measure lines), or missed repetition instructions. While the time signature is of little importance for sheet music-audio synchronization, the other error types can have a strong impact on the alignment result. To achieve high quality alignments, these kinds of errors should be corrected, either by offering user interfaces for manual intervention or by developing new OMR techniques improving on those specific deficits.

Another shortcoming of most OMR systems is the interpretation of textual information in the score. While some systems are capable of determining text such as lyrics correctly, text-based instructions on dynamics, title headings, and instruments are often recognized without associating their (musical) meaning or are not detected at all. For sheet music-audio synchronization the most significant textual information is the one on transposing instruments.[16] If transposing instruments are part of the orchestra and their specific transposition is not considered during the reconstruction, their voices will be shifted with respect to the remaining score, see Figure 7. However, to the best of our knowledge, no OMR system considers this type of information and attempts its detection.

[16] For transposing instruments, the sounding pitches are several semitones higher/lower than the notes written in the score.

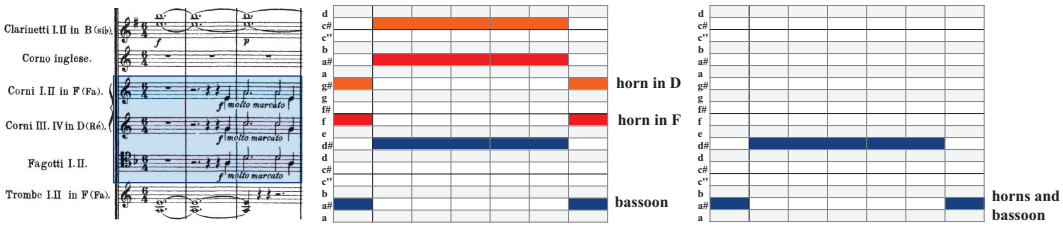


Figure 7 Voices of transposing instruments are shifted with respect to other voices if their transpositions are not known. **Middle:** Erroneous reconstruction in absence of transposition information. **Right:** Correct symbolic representation of the highlighted score extract.

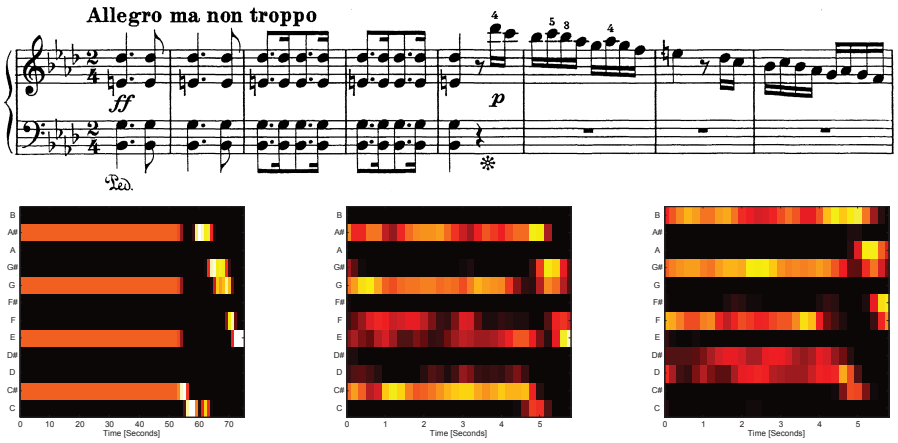


Figure 8 Illustration of chroma features for the first few measures from the third movement of Beethoven's *Piano Sonata No. 23*. The color values represent the intensity of a chroma at a given position (black: low intensity, red: medium intensity and yellow/white: high intensity). The left diagram shows a chroma sequence created from the depicted sheet music extract. The middle and the right diagram show the chroma features for two audio interpretations of the same music extract. The chroma features clearly capture the higher tuning (by one semitone) of the second recordings.

4 Sheet Music-Audio Synchronization

The goal of sheet music-audio synchronization is to link regions in two-dimensional score images to semantically corresponding temporal sections in audio recordings. Therefore, the two data sources need to be made comparable by transforming them into a common mid-level representation. In the synchronization context, chroma-based music features turned out to be a powerful and robust mid-level representation [7, 31]. A chroma vector represents the energy distribution among the 12 pitch classes of the equal-tempered chromatic scale (C, C$^{\#}$, D, ..., B) for a given temporal section of the data, see Figure 8. Chroma features have the property of eliminating differences in timbre and loudness to a certain extend while preserving the harmonic progression in the music. Therefore, their application is most reasonable for music with a clear harmonic progression, like most Western classical music. In addition, by choosing the size of the sections represented by individual chroma vectors appropriately, local errors in the input data can be canceled out for the most part. To transform sheet music into chroma, OMR is performed on the score scans. Afterwards, a MIDI file is created from this data assuming a fixed tempo and standard tuning (see [34] for more information).

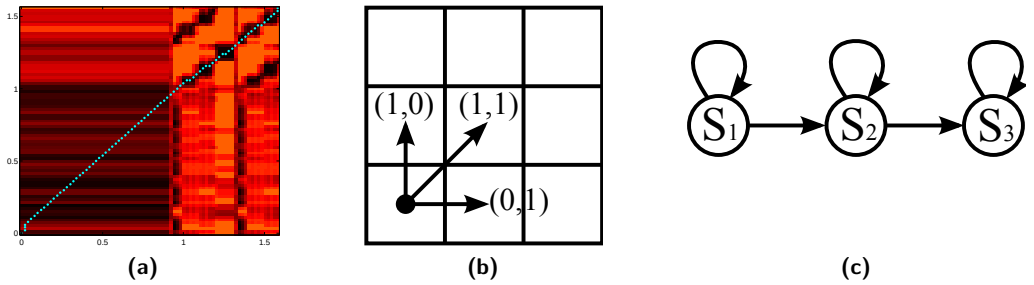


Figure 9 **(a)** Local cost matrix for the score chromagram and one of the audio chromagrams depicted in Figure 8. The optimal alignment path is highlighted in light blue. **(b)** Example of allowed steps during DTW based synchronization. **(c)** Illustration of a left-to-right connected HMM with 3 states.

At the moment, we assume that we are given one sheet music representation and one audio interpretation of the same piece of music. We will address the issue of sheet music-audio mapping in Section 4.1. Furthermore, we will for now assume that the structure of the score and the audio recording coincide. Some ideas on how to handle structural differences will be presented in Section 4.2. After calculating chroma features for both music representations, a local cost matrix can be constructed by pair-wise measuring the similarity between the vectors of the two chroma sequences. Then, the goal is the identification of a path through this matrix that is connecting the two beginnings and endings of the feature sequences and is optimal with respect to the local costs along the path (*optimal alignment path*). See Figure 9a for an example.

There exist two commonly used computational approaches to this task. The first approach is called *dynamic time warping* (DTW) and is based on dynamic programming techniques [1, 18, 24, 31, 42, 43]. After creating the local cost matrix using an appropriate cost measure an accumulated cost matrix is constructed. In this matrix the entry at position (n, m) contains the minimal cost of any alignment path starting at $(1, 1)$ and ending at (n, m). However, during the creation of the alignment path only a certain set of steps is allowed to move through the matrix, e.g., $\{(1, 0), (0, 1), (1, 1)\}$, see Figure 9b. The optimal alignment path is then constructed by backtracking through the matrix using the allowed steps. At each point we chose the predecessor with the lowest accumulated costs. The second approach applies *Hidden Markov Models* (HMM) to determine the optimal alignment path [36, 37, 46, 48]. In this scenario one of the feature sequences is used as hidden states of the HMM and the other sequence forms the set of observations. Usually a left-to-right connected HMM structure is used for score-audio synchronization, see Figure 9c.

In combination with chroma features these alignment techniques allow for some variations in timbre, loudness, and tempo. In addition, small deviations in the data streams (due to errors) can be handled. In contrast, tuning differences are not considered by the presented approaches. Here, the feature sequences show significant differences that can result in a poor synchronization quality (see Figure 8). To suitably adjust the chroma features, a tuning estimation step can be included in the feature calculation process [19]. Instead, one may also apply brute-force techniques such as trying out all possible cyclic shifts of the chroma features [30, 39]. Thus, the presented approaches already cope with some of the problems mentioned in Section 2. In the remainder of this section we want to introduce approaches tackling some of the remaining unsolved problems (i.e., structural differences, certain types of errors, and sheet music-audio mapping).

4.1 Sheet Music-Audio Mapping

Arranging the music data in a digital library in a work-centered way or, more precisely, piece of music-wise has proven beneficial. Thus in the context of a digitization project to build up a large digital music library, one important task is to group all documents that belong to the same piece of music, see Figure 1. Note that in this scenario, the music documents that are to be organized are not given as individual songs or movements, but rather as complete sheet music books or audio CD collections that usually contain several pieces of music.[17] In addition, we typically have to deal with numerous versions of audio recordings of one and the same piece of music,[18] and also with a number of different score versions (different publishers, piano reductions, orchestra parts, transcriptions, etc.) of that piece. Thus, the final goal at this level of detail is to segment both the score books and the audio recordings in such a way that each segment corresponds to one piece of music. Furthermore, each segment should be provided with the appropriate metadata. This segmentation and annotation process, called *sheet music-audio mapping*, is a crucial prerequisite for the sheet music-audio synchronization described in the previous section. One possibility to solve this task is to manually perform this segmentation and annotation. However, for large collections this would be an endless undertaking. Thus semi-automatic or even fully automatic mapping techniques should be developed.

For audio recordings and short audio extracts, music identification services like *Shazam*[19] can provide a user with metadata. Furthermore, ID3 tags, CD covers, or annotation databases such as Gracenote[20] and DE-PARCON[21] can contain information on the recorded piece of music. However, their automated interpretation can quickly become a challenging task. To name just two prominent issues, the opus numbers given by the different sources might not use the same catalogue or the titles might be given in different spellings or different languages. Furthermore, the mentioned services do not provide information for public domain recordings. Another issue can be introduced by audio tracks containing several pieces of music. Here, the exact start and end positions of the individual pieces of music have to be determined.[22] However, this information is usually not provided on CD covers or in metadata databases. Still, the mentioned information sources can be used to support the manual segmentation and annotation process. The automatic extraction and analysis of textual information on scanned score images has to be considered at least equally challenging.

Given one annotated audio recording of all the pieces contained in a score book, Fremerey et al. [25, 27] propose an automatic identification and annotation approach for sheet music that is based on content-based matching. One key strategy of the proposed procedure is to reduce the two different types of music data, the audio recordings as well as the scanned sheet music, to sequences of chroma features, which then allow for a direct comparison across the two domains using a variant of efficient index-based audio matching, see [33]. To this end, the scan feature sequence is compared to the audio feature sequence using subsequence dynamic time warping. The resulting matching curve combined with the information on the

[17] In the context of the PROBADO project, the Bavarian State Library in Munich digitized more than 900 sheet music books (approx. 72,000 score pages) and about 800 audio CDs.
[18] For example, the British Library Sounds include recordings of about 750 performances of Beethoven String Quartets, as played by 90 ensembles, see `http://sounds.bl.uk/Classical-music/Beethoven`
[19] `http://www.shazam.com`
[20] `www.gracenote.com`
[21] `http://www.de-parcon.de/mid/index.html`
[22] Usually, longer periods of silence can hint at the beginning of a new piece. However, the direction *attacca* resulting in two successive movements played without a pause, can prevent this clue from existing.

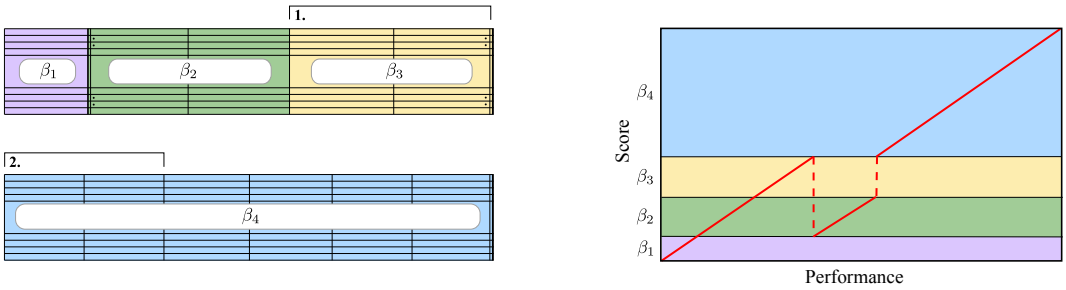


Figure 10 Score block sequence $\beta_1\beta_2\beta_3\beta_4$ created from notated score jumps and alignment path for an audio with block structure $\beta_1\beta_2\beta_3\beta_2\beta_4$ (adapted from [25]).

audio segmentation finally gives both the segmentation and the annotation of the scanned sheet music.

In the same manner, additional audio recordings of already known pieces can be segmented and annotated. Therefore, through the presented approach the manual processing of only one manifestation of each piece of music is required.

4.2 Dealing with Structural Differences

When comparing and synchronizing scores and performances, it may happen that their global musical structures disagree due to repeats and jumps performed differently than suggested in the score. These structural differences have to be resolved to achieve meaningful synchronizations. In the scenario of online score-audio synchronization this issue has already been addressed [2, 35, 44, 51]. Pardo and Birmingham [44] and Arzt et al. [2] both use structural information available in the score data to determine music segments where no jumps can occur. In the first publication an extended HMM is used to allow for jumps between the known segment boundaries. In the second approach an extension of the DTW approach to music synchronization is used to tackle structural differences. At each ending of a section, three hypotheses are pursued in parallel. First, the performance continues on to the next section. Second, the current section is repeated. Third, the subsequent section is skipped. After enough time has passed in the performance the most likely hypothesis is kept and followed. Besides approaches exploiting structural information available from the score, Müller et al. [38, 40] approached a more general case where two data sources (e.g., two audio recordings) are given but no information on allowed repeats or jumps is available. In this case, only partial alignments of possibly large portions of the two documents to be synchronized are computed.

Fremerey et al. [25, 26] presented a method for offline sheet music-audio synchronization in the presence of structural differences, called *JumpDTW*. Here, jump information is derived from the sheet music reconstruction thus creating a block segmentation of the piece of music (see Figure 10). As already mentioned, OMR systems may not recognize all types of jump instructions (especially, textual instructions are often missed). Therefore, bold double bar lines are used as block boundaries. At the end of each block the performance can then either continue to the next block or jump to the beginning of any other block in the piece, including the current one (in contrast to [2] where only forward jumps skipping at most one block are considered). To allow for jumps at block endings, the set of DTW steps is modified. For all block endings, transitions to all block starts in the score are added to the usual steps. By calculating an optimal alignment path using a thus modified accumulated cost matrix,

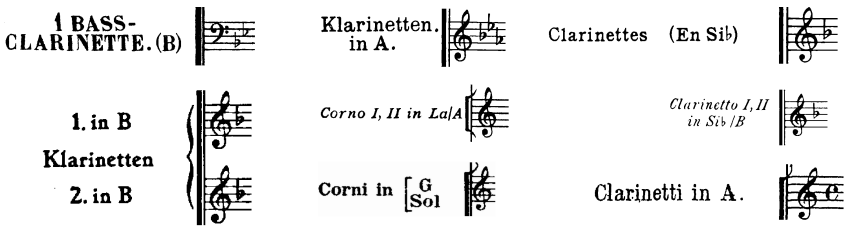


Figure 11 Examples of transposition labels applied by different editors.

possible jumps in the performance can be detected and considered during the synchronization process.

4.3 Dealing with Orchestral Music

Because of the large number of instruments in orchestral music, the score notation inevitably becomes more complex. Typically, this results in a decreased OMR accuracy. Furthermore, orchestral scores contain information commonly neglected by OMR systems. One very important example is the transposition information. The specific transposition of an instrument is usually marked in the score by textual information such as "Clarinet in E", see Figure 11. Obviously, by disregarding this information during the OMR reconstruction, the pitch information for transposing instruments will be incorrect. In the context of sheet music-audio synchronization such global errors in the reconstructed symbolic score data can result in a significant accuracy loss [52, 53].

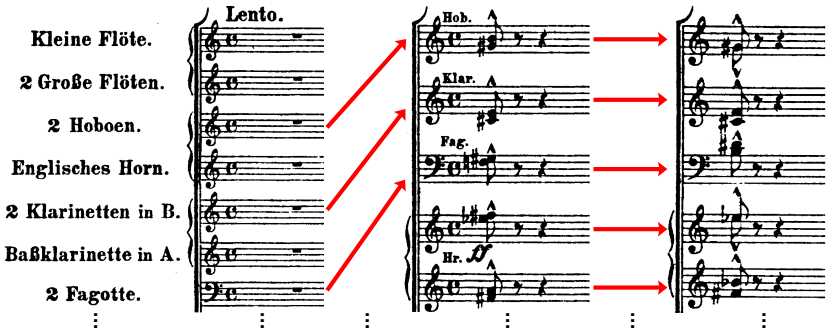


Figure 12 Extracts from Franz Liszt: *Eine Sinfonie nach Dantes Divina Commedia* using compressed notation (publisher: *Breitkopf & Härtel*).

In Western classical music, the score notation usually obeys some common typesetting conventions. Examples are the textual transposition information but also the introduction of all instruments playing in a piece of music by labeling the staves of the first system. Furthermore, a fixed instrument order and the usage of braces and accolades help in reading the score [49]. But despite of all these rules, the task of determining which instrument is supposed to play in a given staff (instrument-staff mapping) and whether or not it is a transposing instrument can be challenging. For most scores the number of staves remains constant throughout the entire piece of music. Therefore the instrument names and transposition information are often omitted after the first system and the information given in the first system needs to be passed on to the remaining systems. The task of determining the instrument of a staff and its transposition becomes even more complicated for compressed score notations where staves of pausing instruments are removed (see Figure 12). Here, the

instrument order is still valid, but some of the instruments introduced in the first system may be missing. To clarify the instrument-staff mapping in these cases, textual information is given. However, in these cases the instrument names are usually abbreviated and therefore more difficult to recognize. Furthermore, transposition information is often only provided in the first system of a piece or in the case that the transposition changes. The textual information might be omitted altogether if the instrument-staff mapping is obvious for a human reader (e.g., strings are always the last instrument group in a system).

Although a great deal of research on OMR has been conducted (see, e.g., [4, 29]), the particular challenges of orchestral scores have not yet been addressed properly. A first approach for the reconstruction of the transposition information was presented in [53]. The instrument-staff mapping as well as the transposition information are reconstructed during three distinct processing steps.

In the first step, the textual information available on the score scans is recovered and interpreted to regain as many instrument labels and transposition labels as possible, see Figure 13. Using the staff location information available in the OMR result, image regions that possibly contain text/words naming an instrument or a transposition are detected and processed by an OCR engine. Subsequently, the detected instruments are mapped to the according staves. To account for different spellings and abbreviations, a library of all possible textual representations of the instruments is used as additional knowledge. Transpositions are recognized by searching for the keyword "in" followed by a valid transposition information.

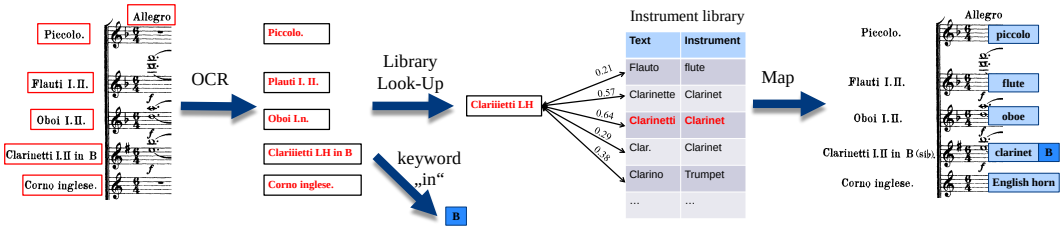


Figure 13 Overview: Reconstruction of instrument and transposition labels from the textual information in the score.

In the second step, the reconstruction from the previous step is used as initialization of an iterative process, see Figure 14. To this end, musical knowledge and common notation conventions are employed. As both the OCR-reconstruction and all information deduced through musical knowledge are uncertain, all instrument-staff mappings are equipped with plausibility values. Besides filling missing mappings, the following iterative update process also strengthens/weakens existing plausibilities. Each iteration of step two can again be divided into three parts. First, the already detected instrument information is successively propagated between consecutive systems by employing the convention that the initially established instrument order is not altered. If two instruments occur in both systems and the number of intermediate staves between these instruments coincides, the instrument information of the intermediate staves of the first system is propagated to the according staves in the subsequent system. Second, musical properties such as "trombone and tuba play in subsequent staves and are grouped by an accolade" are deduced from the score and employed to determine the instrumentation. In the third and final part, the instrument order established in the first system is used again. For all subsequent systems deviations from this order are determined and the according instrument-staff mappings are weakened.

In the last step of the proposed method, the transposition labels given in the first system (and reconstructed in step one) are transferred to the remaining systems. Thereby a

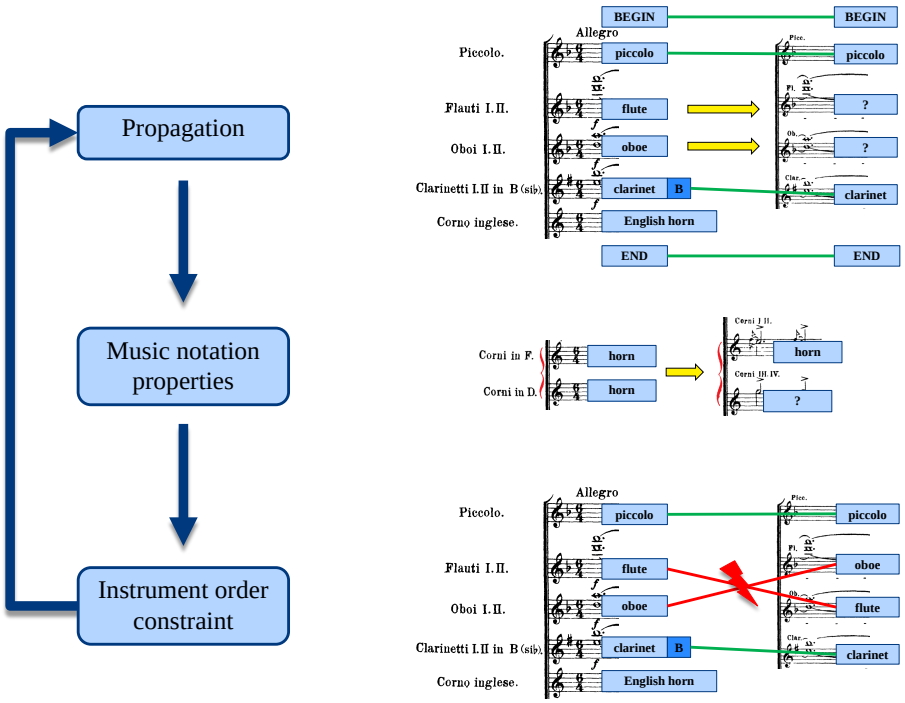


Figure 14 Overview: Recursive approach to the reconstruction of missing instrument and transposition labels.

global correction of the transposition information is achieved even if textual transposition information is only available in the first system.

5 Applications of Sheet Music-Audio Synchronization

In Section 1 we already touched upon possible applications of sheet music-audio synchronization. In this section we first give a more detailed overview of existing user interfaces that employ synchronization techniques (using sheet music or symbolic score data). Then, we focus on current issues in *music information retrieval* (MIR) and show how to incorporate sheet music-audio synchronization to solve specific MIR tasks.

5.1 User Interfaces

The *Laboratorio di Informatica Musicale* at the University of Milan developed the IEEE 1599 standard for the comprehensive description of music content. The proposed XML-format can handle and relate information of various kinds including music symbols, printed scores, audio recordings, text, and images. In addition, music analysis results and synchronization information can be stored as well. Based on this IEEE standard, user interfaces for the simultaneous presentation of multiple music documents have been proposed [3, 5, 6]. To this end, the synchronization results are used for enhanced, multimodal music navigation. At the moment, the synchronization information is created manually but work towards automated score-audio synchronization has been reported [14]. Another project that uses manually

created alignment information is the *Variations* project [21].[23] The goal of *Variations* is the development of a digital music library system to be used in the education context. The system offers music analysis and annotation tools (e.g., structure analysis, time stretching) and page-wise score-audio synchronization. Work on automated synchronization has been described in [47].

WEDELMUSIC is one of the first systems presenting sheet music and audio data simultaneously [8]. During playback a marker moves through the sheet music to identify the currently audible musical position. In addition, page turning is performed automatically by gradually replacing the current sheet/system with the next one. However, the employed automatic synchronization approach was rather simple. Using the start and end points in the sheet music and the audio as anchor points, linear interpolation was applied. As local tempo deviations may result in alignment errors, a user interface for the manual rework of the proposed synchronization was available. Xia et al. [55] present a rehearsal management tool for musicians that exploits semi-automated score-audio synchronization. Here, recordings of various rehearsals are clustered and aligned to a score representation of the piece of music. Additional challenges are introduced by the fact that the recordings can differ in length and may cover different parts of the piece. In the PROBADO project, a digital music library system for the management of large document collections was developed (see Figure 15). The most prominent features are content-based retrieval techniques and a multimodal music presentation implemented by sheet music-audio synchronization [15, 16]. The alignment structures are calculated nearly automatically in this system.

Another application designed to support musicians is automated accompaniment. To this end, online score-audio synchronization determines the current position in the score as well as the current tempo to replay a time-stretched audio recording. Two well known accompaniment systems are *Music Plus One* by Raphael [46, 48] and ANTESCOFO by Cont [13].

5.2 MIR Research

There are various MIR tasks that exploit score information as additional knowledge. For example, in score-informed source separation one assumes that along with the audio recording a synchronized MIDI file is given. Through this file the occurring note events along with their position and duration in the audio are specified. We refer to Ewert and Müller [23] for an extensive overview. At the moment, all approaches use symbolic score data (e.g., MIDI) but sheet music may be applicable as well. However, in this case recognition errors need to be considered by the source separation method. A similar task is the estimation of note intensities in an audio recording where the notes are specified by a symbolic representation [22]. Again, to avoid the manual creation of a MIDI file, the exploitation of score scans together with sheet music-audio synchronization techniques, seems reasonable.

Another important research topic is lyrics-audio synchronization [28, 32]. Instead of using the commonly employed speech analysis techniques, sheet music can be added as additional information. Thereby, the lyrics can be derived from the OMR results. Afterwards, the lyrics-audio alignment can be calculated by means of the sheet music-audio synchronization [15, 41, 50].

[23] http://www.dlib.indiana.edu/projects/variations3

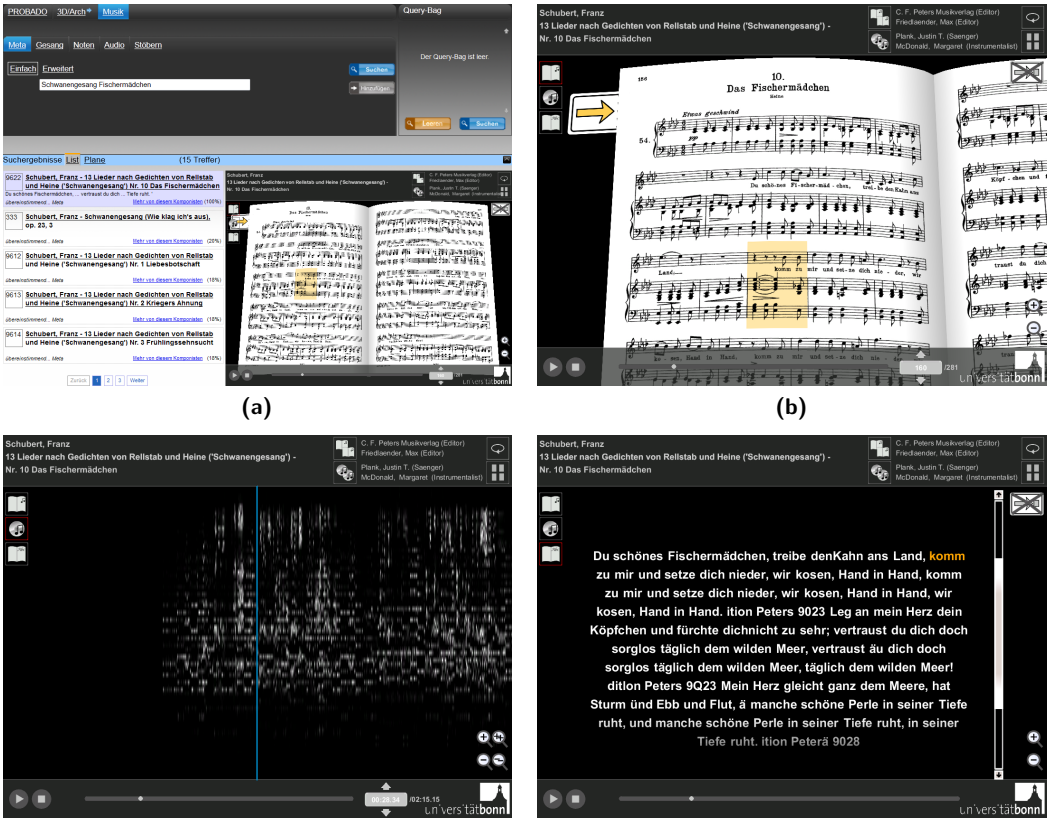


Figure 15 The PROBADO music user interface. **(a)** Search interface with a result presentation on the piece of music level. On the bottom right, access to all documents containing the selected piece is provided. Besides sheet music **(b)**, the interface offers visualizations of audio recordings **(c)** and lyrics **(d)**. Sheet music-audio synchronization results allow for the currently audible measure to be highlighted. Equally, the sung word is marked in the lyrics [50]. Different sheet music edition or other audio recordings can easily be selected. During a document change, the linking structures help to preserve the musical position and playback continues smoothly.

There are several other tasks where score-audio synchronization might help reducing the complexity of the problem. Some examples are structure analysis, chord recognition, and melody extraction.

6 Outlook

Although, current sheet music-audio synchronization algorithms perform quite well, there still exist some open issues. First, to allow for a higher level of detail, the input data has to become more reliable. In particular, the OMR accuracy needs to be improved. After achieving a high-resolution synchronization, e.g., on the note level, the question of how to present this alignment structure arises. For orchestral music, highlighting the currently audible notes in all voices would result in a very nervous visualization. At the moment only printed sheet music of reasonable quality is being used. However, huge amounts of old sheet music volumes exist that are heavily yellowed and stained. In addition, large collections of handwritten scores and hand-annotated printed sheet music are available. Some OMR

systems are capable of dealing with those types of sheet music but the applicability of the resulting symbolic representation (in terms of recognition accuracy) to the synchronization task would have to be investigated.

In Section 4.1, we discussed the task of sheet music-audio mapping and presented a method for segmenting and identifying score data using already segmented and identified audio documents. With this approach, at least one version of a piece of music has to be manually annotated. For large music databases a full automation or at least some support in the unavoidable manual tasks is highly desired. Looking at sheet music and CD booklets, they contain a wealth of textual information (composer, title, opus number, etc.). Automatically detecting and interpreting this information constitutes an important future step.

One can think of a variety of applications that would benefit from the presented synchronization techniques. The method could be extended to allow for online score following and live accompaniment of musicians using a scanned score. In [20, 44] the synchronization of lead sheets with a fully instrumented audio recording was suggested. In a similar manner, the sheet music of individual voices could be synchronized to an orchestra recording. These linking structures could for example be of use in the context of digital orchestra stands [9]. All parts are synchronized to the conductors score and upon selecting a position in the conductors score the position in all score visualization changes accordingly.

7 Acknowledgment

This work was supported by the German Research Foundation DFG (grants CL 64/7-2 and CL 64/6-2). Meinard Müller is funded by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI). We would like to express our gratitude to Maarten Grachten, Masataka Goto, and Markus Schedl for their helpful and constructive feedback.

References

1 Vlora Arifi, Michael Clausen, Frank Kurth, and Meinard Müller. Automatic synchronization of musical data: A mathematical approach. *Computing in Musicology*, 13:9–33, 2004.

2 Andreas Arzt, Gerhard Widmer, and Simon Dixon. Automatic page turning for musicians via real-time machine listening. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 241–245, Patras, Greece, 2008.

3 Denis Baggi, Adriano Baratè, Goffredo Haus, and Luca Andrea Ludovico. NINA—navigating and interacting with notation and audio. In *Proceedings of the International Workshop on Semantic Media Adaptation and Personalization (SMAP)*, pages 134–139, Washington, DC, USA, 2007. IEEE Computer Society.

4 David Bainbridge and Tim Bell. The challenge of optical music recognition. *Computers and the Humanities*, 35(2):95–121, 2001.

5 Adriano Baratè, Goffredo Haus, and Luca A. Ludovico. IEEE 1599: a new standard for music education. In *Proceedings of the International Conference on Electronic Publishing (ELPUB)*, pages 29–45, Milan, Italy, 2009.

6 Adriano Baratè, Luca A. Ludovico, and Alberto Pinto. An IEEE 1599-based interface for score analysis. In *Computer Music Modeling and Retrieval (CMMR)*, Copenhagen, Denmark, 2008.

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

8 Pierfrancesco Bellini, Ivan Bruno, Paolo Nesi, and Marius B. Spinu. Execution and synchronisation of music score pages and real performance audios. In *Proceedings of the IEEE Inter-*

national Conference on Multimedia and Expo (ICME), pages 125–128, Lausanne, Switzerland, 2002.

9 Pierfrancesco Bellini, Fabrizio Fioravanti, and Paolo Nesi. Managing music in orchestras. *Computer*, 32:26–34, 1999.

10 Esben Paul Bugge, Kim Lundsteen Juncher, Brian Søborg Mathiasen, and Jakob Grue Simonsen. Using sequence alignment and voting to improve optical music recognition from multiple recognizers. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 405–410, 2011.

11 Donald Byrd, William Guerin, Megan Schindele, and Ian Knopke. OMR evaluation and prospects for improved OMR via multiple recognizers. Technical report, Indiana University, Bloomington, Indiana, USA, 2010.

12 Donald Byrd and Megan Schindele. Prospects for improving OMR with multiple recognizers. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 41–46, Victoria, Canada, 2006.

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

14 Antonello D'Aguanno and Giancarlo Vercellesi. Automatic music synchronization using partial score representation based on IEEE 1599. *Journal of Multimedia*, 4(1):19–24, 2009.

15 David Damm. *A Digital Library Framework for Heterogeneous Music Collections—from Document Acquisition to Cross-modal Interaction*. PhD thesis, University of Bonn (in preparation), 2012.

16 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: Special Issue on Music Digital Libraries (to appear)*, 2012.

17 Roger B. Dannenberg. An on-line algorithm for real-time accompaniment. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 193–198, 1984.

18 Johanna Devaney, Michael I. Mandel, and Daniel P. W. Ellis. Improving MIDI-audio alignment with acoustic features. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 45–48, New Paltz, NY, USA, 2009.

19 Karin Dressler and Sebastian Streich. Tuning frequency estimation using circular statistics. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 357–360, Vienna, Austria, 2007.

20 Zhiyao Duan and Bryan Pardo. Aligning semi-improvised music audio with its lead sheet. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 513–518, Miami, FL, USA, 2011.

21 Jon W. Dunn, Donald Byrd, Mark Notess, Jenn Riley, and Ryan Scherle. Variations2: Retrieving and using music in an academic setting. *Communications of the ACM, Special Issue: Music information retrieval*, 49(8):53–48, 2006.

22 Sebastian Ewert and Meinard Müller. Estimating note intensities in music recordings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 385–388, Prague, Czech Republic, 2011.

23 Sebastian Ewert and Meinard Müller. Score-informed source separation. In Meinard Müller, Masataka Goto, and Markus Schedl, editors, *Multimodal Music Processing (to appear)*, Dagstuhl Follow-Ups. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2012.

24 Sebastian Ewert, Meinard Müller, and Peter Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of the IEEE International Conference*

on Acoustics, Speech, and Signal Processing (ICASSP), pages 1869–1872, Taipei, Taiwan, 2009.

25 Christian Fremerey. *Automatic Organization of Digital Music Documents – Sheet Music and Audio*. PhD thesis, University of Bonn, 2010.

26 Christian Fremerey, Meinard Müller, and Michael Clausen. Handling repeats and jumps in score-performance synchronization. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 243–248, Utrecht, The Netherlands, 2010.

27 Christian Fremerey, Meinard Müller, Frank Kurth, and Michael Clausen. Automatic mapping of scanned sheet music to audio recordings. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 413–418, Philadelphia, USA, 2008.

28 Hiromasa Fujihara, Masataka Goto, Jun Ogata, and Hiroshi G. Okuno. Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1252 –1261, 2011.

29 Ichiro Fujinaga. Optical music recognition bibliography. `http://ddmal.music.mcgill.ca/wiki/Optical_Music_Recognition_Bibliography`, 2000.

30 Masataka Goto. A chorus-section detecting method for musical audio signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 437–440, Hong Kong, China, 2003.

31 Ning Hu, Roger B. Dannenberg, and George Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, US, 2003.

32 Min-Yen Kan, Ye Wang, Denny Iskandar, Tin Lay New, and Arun Shenoy. LyricAlly: Automatic synchronization of textual lyrics to acoustic music signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):338 –349, 2008.

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

34 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 (ISMIR)*, pages 261–266, Vienna, Austria, 2007.

35 John Lawter and Barry Moon. Score following in open form compositions. In *Proceedings of the International Computer Music Conference (ICMC)*, Ann Arbor, MI, USA, 1998.

36 Nicola Montecchio and Arshia Cont. A unified approach to real time audio-to-score and audio-to-audio alignment using sequential montecarlo inference techniques. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 193–196, Prague, Czech Republic, 2011.

37 Nicola Montecchio and Nicola Orio. A discrete filter bank approach to audio to score matching for polyphonic music. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 495–500, Kobe, Japan, 2009.

38 Meinard Müller and Daniel Appelt. Path-constrained partial music synchronization. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 65–68, Las Vegas, Nevada, USA, 2008.

39 Meinard Müller and Michael Clausen. Transposition-invariant self-similarity matrices. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 47–50, Vienna, Austria, 2007.

40 Meinard Müller and Sebastian Ewert. Joint structure analysis with applications to music annotation and synchronization. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 389–394, Philadelphia, Pennsylvania, USA, 2008.

41 Meinard Müller, Frank Kurth, David Damm, Christian Fremerey, and Michael Clausen. Lyrics-based audio retrieval and multimodal navigation in music collections. In *Proceedings of the European Conference on Digital Libraries (ECDL)*, pages 112–123, Budapest, Hungary, 2007.

42 Bernhard Niedermayer. Improving accuracy of polyphonic music-to-score alignment. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 585–590, Kobe, Japan, 2009.

43 Nicola Orio and François Déchelle. Score following using spectral analysis and hidden Markov models. In *Proceedings of the International Computer Music Conference (ICMC)*, Havanna, Cuba, 2001.

44 Bryan Pardo and William P. Birmingham. Modeling form for on-line following of musical performances. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 1018–1023, Pittsburgh, PA, USA, 2005.

45 Dennis Howard Pruslin. *Automatic recognition of sheet music*. PhD thesis, Massachusetts Institute of Technology, 1966.

46 Christopher Raphael. Music Plus One: A system for flexible and expressive musical accompaniment. In *Proceedings of the International Computer Music Conference (ICMC)*, Havana, Cuba, 2001.

47 Christopher Raphael. A hybrid graphical model for aligning polyphonic audio with musical scores. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 387–394, Barcelona, Spain, 2004.

48 Christopher Raphael. Music Plus One and machine learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, Haifa, Israel, 2010.

49 Stanley Sadie, editor. *The New Grove Dictionary of Music and Musicians (second edition)*. Macmillan, London, 2001.

50 Markus Schäfer. Hochwertige automatische Extraktion von Gesangstext aus Notenbänden und mediensynchrone Darstellung. Diploma thesis, University of Bonn (in preparation), 2012.

51 Mevlut Evren Tekin, Christina Anagnostopoulou, and Yo Tomita. Towards an intelligent score following system: Handling of mistakes and jumps encountered during piano practicing. In *Computer Music Modeling and Retrieval (CMMR)*, pages 211–219, Pisa, Italy, 2005.

52 Verena Thomas, Christian Fremerey, Sebastian Ewert, and Michael Clausen. Notenschrift-Audio Synchronisation komplexer Orchesterwerke mittels Klavierauszug. In *Proceedings of the Deutsche Jahrestagung für Akustik (DAGA)*, pages 191–192, Berlin, Germany, 2010.

53 Verena Thomas, Christian Wagner, and Michael Clausen. OCR-based post-processing of OMR for the recovery of transposing instruments in complex orchestral scores. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 411–416, Miami, FL, USA, 2011.

54 Barry Vercoe. The synthetic performer in the context of live performance. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 199–200, 1984.

55 Guangyu Xia, Dawen Liang, Roger B. Dannenberg, and Mark J. Harvilla. Segmentation, clustering, and display in a personal audio database for musicians. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 139–144, Miami, FL, USA, 2011.

Lyrics-to-Audio Alignment and its Application

Hiromasa Fujihara and Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
{h.fujihara, m.goto}@aist.go.jp

Abstract

Automatic lyrics-to-audio alignment techniques have been drawing attention in the last years and various studies have been made in this field. The objective of lyrics-to-audio alignment is to estimate a temporal relationship between lyrics and musical audio signals and can be applied to various applications such as Karaoke-style lyrics display. In this contribution, we provide an overview of recent development in this research topic, where we put a particular focus on categorization of various methods and on applications.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, H.5.1 Multimedia Information Systems

Keywords and phrases Lyrics, Alignment, Karaoke, Multifunctional music player, Lyrics-based music retrieval

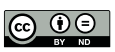
Digital Object Identifier 10.4230/DFU.Vol3.11041.23

1 Introduction

Music is an important media content in both industrial and cultural aspects, and a singing voice (vocal) is one of the most important elements of music in many music genres, especially in popular music. Thus, research that deals with singing voices is gaining in importance from cultural, industrial and academic perspectives. Lyrics are one of the most important aspects of singing voices. Since the lyrics of a song represent its theme and story, they are essential for creating an impression of the song. When a song is heard, for example, most people would follow the lyrics while listening to the vocal melody. This is why music videos often help people to enjoy music by displaying synchronized lyrics as a Karaoke-style caption.

In this paper we overview several research attempts that deal with automatic synchronization between music and lyrics, also known as lyrics-to-audio alignment. To deal with lyrics in music, one of the ultimate goals is automatic lyric recognition (i.e., the dictation of lyrics in a mixture of singing voices and accompaniments). However, since this goal has not yet been achieved even for ordinary speech in noisy environments with satisfactory accuracy, it is not a realistic way to pursue automatic dictation of the lyrics in the first place. As a matter of fact, though several research attempts have been made to pursue this goal [23, 19, 21, 5, 14], none of them achieved satisfactory performance in natural environments under realistic assumptions so far. From this perspective, it can be said that lyrics-to-audio alignment is a reasonable problem setting because not only does the problem itself have a number of practical applications but knowledge accumulated by tackling this problem can also be a stepping-stone for automatic lyric recognition.

The rest of this paper is organized as follows. We continue with defining the problem of lyrics-to-audio alignment and describing main challenges of this task. Then, Section 3 summarizes numerous study attempts and introduces some representative works. In Section 4, we introduce applications of lyrics-to-audio alignment techniques.



Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 23–36

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

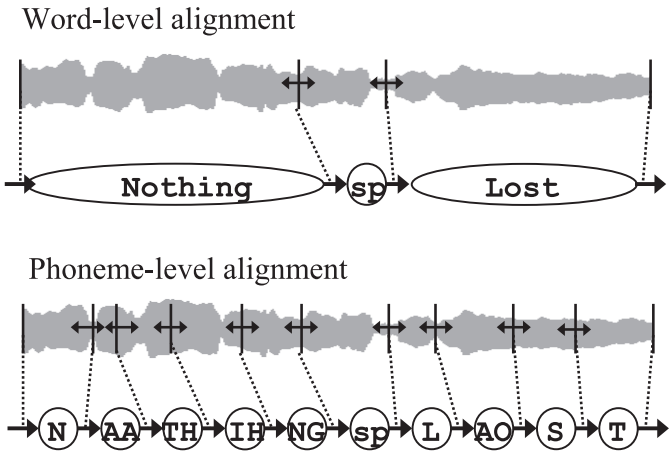


Figure 1 Example of word-level alignment and phoneme-level alignment.

2 Lyrics-to-Audio Alignment

2.1 Problem Definition and Applications

Given audio signals of singing voices and corresponding textual lyrics as input data, lyrics-to-audio alignment can be defined as a problem of estimating the temporal relationship between them. To this end, start and end times of every block of certain length in lyrics are estimated. Here, the term "block" means a fragment of lyrics, the size of which depends on the application as described below, and can be either phoneme, syllable, word, phrase, line, or paragraph (See Figure 1).

Numerous applications of this technique are conceivable, such as a music player with Karaoke-like lyrics display function, automatic generation of subtitles of music videos, and the generation of audio thumbnails. Apart from these consumer-oriented applications, this technique can also be used as a basic building block for other singing voice research, such as singing voice synthesis [16] and analysis of the relationship between musical audio signals and lyrics [17]. In the case of music video subtitles, granularity of synchronization does not have to be very precise and line or phrase level alignment is sufficient. If the precise timing of the lyrics is needed such as in the case of Karaoke-like display, on the other hand, phoneme or word level alignment is imperative.

2.2 Difficulties

The problem of lyrics-to-audio alignment bears a relationship to text-to-speech alignment used in automatic speech recognition research, which is generally conducted by using a forced alignment techniques with mel-frequency cepstral coefficients (MFCCs), phonetic hidden Markov models (HMMs), and the Viterbi algorithm[1] [25]. However, it is difficult to directly apply the forced alignment technique to singing voices because there are several difficulties

[1] Hereafter, we use a term "forced alignment" to refer to this particular technique that can align transcribed text and speech signals by using phonetic HMMs and the Viterbi algorithm.

intrinsic to singing voices:

1. **Fluctuation of acoustic characteristics.** It is known that the singing voice has more complicated frequency and dynamic characteristics than speech [20]. For example, fluctuation of fundamental frequency (F0) [2] and loudness of singing voices are far stronger than those of speech sounds.
2. **Influences of accompaniment sounds.** Singing voice signals are generally accompanied by other instruments, which make it difficult even for a human to understand what is being sung. This is mainly because spectrum of the singing voices are overlapped and distorted by those of accompaniment sounds. Thus, it is necessary to either reduce such negative influences or use features robust to them.
3. **Incomplete lyrics.** Available textual lyrics do not always correspond exactly to what is sung in a song. For example, repetitive paragraphs are sometimes omitted for the sake of simplicity and utterances of interjections (such as "yeah" and "oh") are often excluded in the lyrics. This is particularly problematic when lyrics are taken from the Internet [8].

3 Literature Review

After overviewing studies of this field, this section describes brief explanations of representative works.

3.1 Overview of Previous Studies

A number of studies have been made in the field of lyrics-to-audio alignment [10, 6, 24, 15, 9, 13, 7, 3, 12]. Except for early research [10], most of the studies dealt with singing voices in polyphonic popular music. Since lyrics are inevitably language-dependent, it is not easy to prepare training data for a number of several languages. Thus, evaluations were usually conducted by using songs sung in a single language such as English [6, 7], Chinese [24], and Japanese [3]. With that being said, except for a study that specialized in Cantonese [24], most of them are applicable to any language in principle.

These studies can be categorized according to the following two main viewpoints:

1. **Primary cue for aligning music and lyrics.** To achieve an accurate estimation of a temporal relationship between music and lyrics, it is important to deliberately design features (or representation) that are used to represent music and lyrics and methods to compare these features since such features and methods directly affect the performance of an entire system.
2. **Additional methods for improving performance.** Polyphonic audio signals are so complex that it is not easy to align music and lyrics accurately just by using a single method. Thus, many studies have integrated several additional methods and information to improve performance of alignment such as music understanding techniques and musical knowledge.

Table 1 summarizes the conventional studies from these viewpoints.

3.1.1 Primary Cue for Aligning Music and Lyrics

To characterize algorithms for lyrics-to-audio alignment, it is of central importance to categorize what kind of features they extract from audio and lyrics and how they compare

[2] "F0", which represents how high a sound is, is sometimes referred as "pitch" although, strictly speaking, their definitions are different because the F0 is a physical feature while the pitch is a perceptual feature.

Table 1 Summarization of the conventional studies.

Authors	Primary method	Other additional methods
Loscos *et al.* [10]	The forced alignment with MFCCs	
Iskandar *et al.* [6]	The forced alignment with MFCCs	Song structure Musical knowledge
Wong *et al.* [24]	Comparison of F0 contours	Vocal enhancement Vocal detection Onset detection
Müller *et al.* [15]	Audio-to-MIDI alignment with lyrics-enhanced MIDI files	
Lee *et al.* [9]	Dynamic programming with manually-labeled lyrics segmentation	Structural segmentation
Mesaros *et al.* [13]	The forced alignment	Vocal segregation
Kan *et al.* [7]	Phoneme duration	Beat detection Structure detection Vocal detection
Fujihara *et al.* [3]	The forced alignment	Vocal segregation Vocal detection Fricative detection
Mauch *et al.* [12]	The forced alignment with chord labels	Vocal segregation Vocal detection

them. From this viewpoint, conventional studies can be categorized into the following three categories; those that use acoustic phonetic features, those that use other features, and those that use features taken from external sources.

Studies that fall into the first category [10, 6, 13, 3] adopt the forced alignment. It compares phonetic features (such as MFCCs) extracted from audio signals with a phone model consisting of a sequence of phonemes in the lyrics. Since the forced alignment technique is mainly designed for clean speech signals, the main focus of these studies lies in how to apply it to singing voices with accompaniment sounds. For this purpose, most of the studies incorporate various additional methods described in Section 3.1.2.

The second category contains studies that do not use the forced alignment technique. Wong *et al.* [24] used the tonal characteristics of Cantonese language and compared the tone of each word in the lyrics with the F0 of the singing voice. Kan *et al.* developed a system called LyricAlly [7], which used the duration of each phoneme as a main cue for a fine alignment along with structural information for a coarse alignment.

Instead of directly comparing music and lyrics, studies of the third category deploy external information and use them as a cue for alignment. For example, Müller *et al.* [15] used MIDI files that are manually aligned with lyrics. Then, by executing automatic alignment between music recordings and the MIDI file, they indirectly estimated temporal relationship between music and lyrics. Lee *et al.* [9], assuming that manually-annotated segmentation labels (such as Chorus and Verse) are available, aligned these labels with automatically-estimated structural segmentation by using dynamic programming. While these strategies could result in simpler or more accurate alignments, the range of songs to which the algorithms are applicable is inevitably limited.

In addition to the above mentioned works, Mauch *et al.* [12] used both acoustic phonetic

features and external information at the same time. It can be said that this work belongs to both the first and third categories. More specifically, assuming that the textual chord information provided in the paired chords-lyrics format is available, they integrated lyrics-to-audio alignment and chord-to-audio alignment. Chord alignment, which is more reliable but can be done in only bar or note level, worked as a coarse alignment, followed by a fine alignment achieved by singing voice alignment.

3.1.2 Additional Methods for Improving Performance

In addition to the primary cues described above, most of the studies sought to improve their algorithm by incorporating other music understanding and signal processing methods. For example, some studies used vocal detection methods as a preprocessing step [24, 7, 3]. Regions detected as non-vocal are excluded from the allocation of lyrics. Methods for reducing the influence of accompaniment sounds and enhance singing voices were also used in [24, 13, 3]. This process is usually done before extracting features from audio signals to enable feature extractors to accurately capture the characteristics of singing voices. Fujihara *et al.* [3] and Mesaros *et al.* [13] used singing voice segregation techniques based on harmonic structures, and Wong *et al.* [24] used bass and drum reduction and center signal segregation methods. Other information such as beat [7], song structure [9, 7], onset [24], fricative sound [3], and musical knowledge about rhythms and notes [6] were automatically extracted and incorporated.

3.2 A Lyrics-to-Audio Alignment Method for Cantonese Popular Music

Wong *et al.* developed a lyrics-to-audio alignment method based on tonal characteristics of Cantonese popular music [24]. Cantonese, which is a tone language, distinguishes the meaning of a word by changing the pitch level. Their method took advantage of this fact and tried to align music and lyrics by using pitches extracted from audio signals and those estimated from lyrics, assuming that the contour of the lyrics and that of the musical melody match perfectly. In their method, a vocal signal enhancement algorithm based on center signal estimation and bass and drum reduction methods was used to detect the onsets of the syllables and to estimate the corresponding pitches. They then used a dynamic time warping algorithm to align lyrics and music. Figure 2 shows a block diagram of the method.

To estimate the onsets and the pitch accurately, the authors developed a vocal signal enhancement technique. Based on the assumption that only singing voice and drum signals are located at the center in stereo audio signals, they extracted center parts of the stereo recordings by using a spectral subtraction method. Bass and drum sounds were then removed by subtracting the average spectrum within five-second segments.

Then the onsets of vocal notes, which are expected to correspond to syllables, were detected as the smallest unit of alignment. They first extracted the amplitude envelope of the signal and detected candidates of the onsets by using a difference of the amplitude envelope. Finally they eliminated non-vocal onsets by using a neural network classifier with standard audio features such as spectrum flux, zero-crossing rate, and Mel-frequency cepstral coefficients (MFCCs).

As features for aligning music and lyrics, they used pitch contours. Pitch contours of audio signals were extracted by a standard F0 estimation method, and that of lyrics were estimated from the lyrics based on linguistic rules. These two types of pitch contours are aligned by using dynamic time warping.

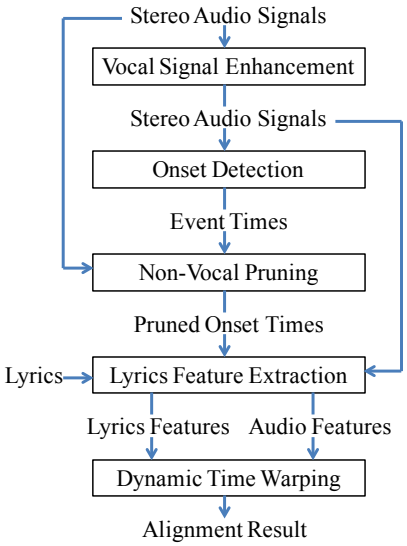


Figure 2 A block diagram of a lyrics-to-audio alignment method in [24].

3.3 LyricAlly

Kan *et al.* developed a lyrics-to-audio alignment system called LyricAlly [7]. It integrates several music understanding techniques such as beat detection, chorus detection, and vocal estimation. They first conducted section-level alignment, which was followed by line-level alignment. Figure 3 shows a block diagram of LyricAlly.

Three kinds of music understanding techniques, namely hierarchical rhythm structure detection, chorus detection, and vocal detection, were executed as a preprocessing step, to constrain and simplify the synchronization process based on musical knowledge. Input lyrics were then analyzed to estimate the duration of the lyrics. This duration estimation process was done based on supervised training.

Assuming a song is consisted of a specific type of song structure (Verse-Chorus-Verse-Chorus) and that each section of lyrics is already marked as a single block, they first conducted section level alignment based on the chorus section detected by using the chorus and vocal detectors. Then, they conducted line-level alignment by using duration information estimated from lyrics.

3.4 A Lyrics-to-Audio Alignment Method based on the forced Alignment

Fujihara *et al.* developed a lyrics-to-audio alignment method based on the forced alignment technique [3]. Because the ordinary forced alignment technique used in automatic speech recognition is negatively influenced by accompaniment sounds performed together with a vocal and also by interlude sections in which the vocal is not performed, they first obtained the waveform of the melody by using a vocal segregation method proposed in [2]. They then detected the vocal region in the separated melody's audio signal, using a vocal detection method based on a Hidden Markov Model (HMM). They also detected the fricative sound and incorporated this information into the next alignment stage. Finally, they aligned the lyrics and the separated vocal audio signals by using the forced alignment technique. Figure

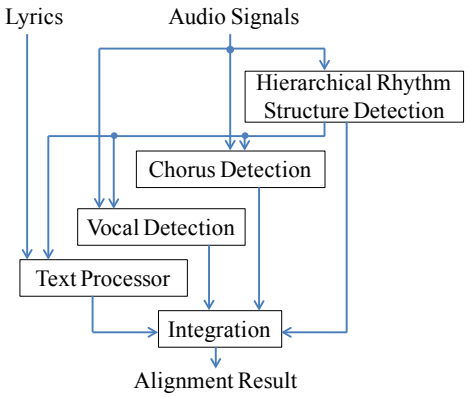


Figure 3 A block diagram of Lyrically [7].

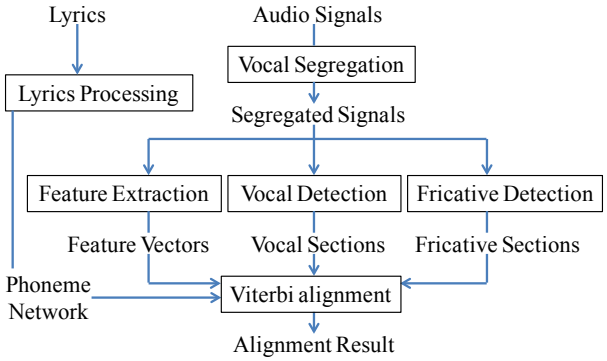


Figure 4 A block diagram of a lyrics-to-audio-alignment method proposed in [3].

4 shows an overview of this method.

Before extracting a feature that represents the phonetic information of a singing voice from polyphonic audio signals, they tried to segregate vocal sound from accompaniment sounds by using a melody detection and resynthesis technique based on a harmonic structure [2]. The technique consists of the following three parts:

1. Estimate the fundamental frequency (F0) of the melody by using a method called PreFEst [4].
2. Extract the harmonic structure corresponding to the melody.
3. Resynthesize the audio signal (waveform) corresponding to the melody by using a sinusoidal synthesis.

This melody resynthesis usually results in vocal signals with bad sound quality for a human perception and it makes it even more difficult for humans to recognize lyrics. However, for a computer, which does not have a sophisticated perceptive system that humans have, this process is important.

The authors developed a vocal detection method to eliminate the influence of non-vocal regions. The method is based on supervised-training of characteristics of singing voices and non-vocal sounds. This method is needed because the melody detection technique assumed that the F0 of the melody is the most predominant in each frame and could not detect

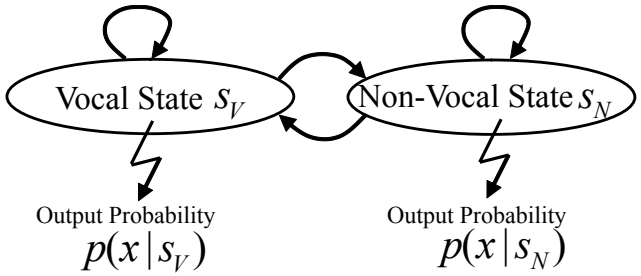


Figure 5 A hidden Markov model (HMM) for vocal activity detection [3].

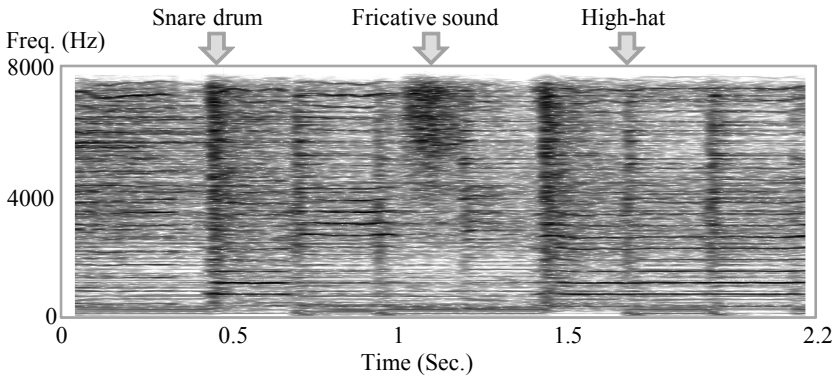


Figure 6 Example spectrogram depicting snare drum, fricative, and high-hat cymbal sounds [3]. The characteristics of fricative sounds are depicted as vertical lines or clouds along the frequency axis, whereas periodic source components tend to have horizontal lines.

regions where vocal does not exist. Thus, such regions have to be eliminated before actually aligning lyrics to segregated signals. An HMM was introduced that transitions back and forth between a vocal state, s_V, and a non-vocal state, s_N, as shown in Figure 5. Vocal state means that vocals are present and non-vocal state means that vocals are absent. Given the feature vectors of input audio signals, x_t, at time t, the problem of vocal detection is finding the most likely sequence of vocal and non-vocal states, $\hat{S} = \{s_1, \cdots, s_t, \cdots\}\,(s_t \in \{s_V, s_N\})$.

$$\hat{S} = \underset{S}{\operatorname{argmax}} \sum_t \{\log p(\boldsymbol{x_t}|s_t) + \log p(s_{t+1}|s_t)\}\,, \tag{1}$$

where $p(\boldsymbol{x}|s)$ represents an output probability of state s, and $p(s_{t+1}|s_t)$ represents a state transition probability for the transition from state s_t to state s_{t+1}. Unlike other previous studies on vocal detection [1, 22, 18], this methods could automatically control the balance between vocal and non-vocal regions.

The forced alignment technique used in automatic speech recognition research synchronizes speech signals and texts by making phoneme networks that consist of all the vowels and consonants. However, since the vocal segregation method, which is based on the harmonic structure of the melody, cannot segregate unvoiced consonants that do not have harmonic structure, it is difficult for the general forced alignment technique to align unvoiced consonants correctly. Therefore, the authors developed a method for detecting unvoiced consonants from the original audio signals. They particularly focused on the unvoiced fricative sounds

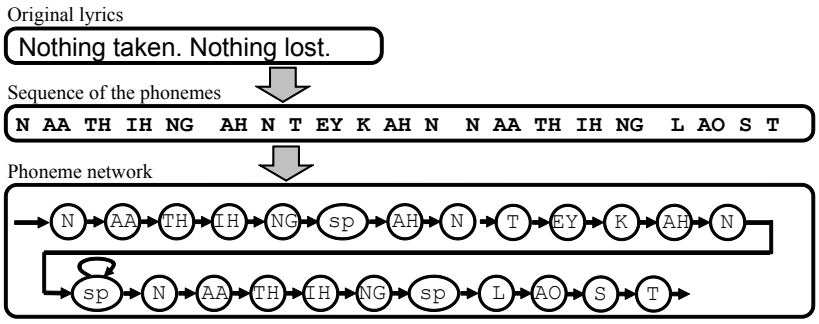


Figure 7 Example for converting from original lyrics to a phoneme network [3].

(a type of unvoiced consonant) because their durations are generally longer than those of the other unvoiced consonants and because they expose salient frequency components in the spectrum. They first suppressed peak components in the spectrum, which is not related to fricative sounds. The fricative sounds were then detected by using the ratio of the power of a band where salient frequency components of fricative sounds exist to that of the other bands. Figure 6 shows an example of a fricative sound to be detected. Then, in the forced alignment stage, fricative consonants were only allowed to appear in the detected candidates of fricative regions.

To actually align lyrics and music, a phoneme network was created from the given lyrics and feature vectors are extracted from separated vocal signals. Figure 7 shows an example of conversion from lyrics to a phoneme network. The phoneme network consists of sequentially connected HMMs of phonemes that appeared in the lyrics. Each HMM represents sound characteristic of a corresponding phoneme and is used to compare likelihood of feature vectors extracted from audio signals. Finally, the forced alignment was executed by calculating the most likely path of a sequence of the feature vectors and the phoneme network. The authors used the proportion of the length of the sections which are correctly labeled as a quality measure and reported that the system achieved 90% accuracy for 8 out of 10 songs.

4 Applications to Music Player and Lyrics-based Music Retrieval

Due to the diffusion of the personal computer and the portable music player, there has been a growing opportunity to listen to songs using devices that have screens. It is natural to consider using that screens to enrich users' experience in music appreciation by displaying lyrics and related information on it. This section introduces three examples of this idea, which display synchronized lyrics estimated by lyrics-to-audio alignment techniques and utilize it to lyrics-based music retrieval or music navigation.

4.1 Lyrics-based Music Retrieval

Müller *et al.* proposed a lyrics search engine based on their lyrics-to-audio alignment [15]. Their system was developed as a plug-in software on the SyncPlayer framework, which is a software framework that integrates various MIR-techniques. Figure 8 (taken from [15]) shows a screenshot of their system. The three windows on the left side display lyrics synchronously and the right window enables lyrics-based search. It should be noted that users can directly

Figure 8 Screenshot of lyric-based search system developed on the SyncPlayer framework [15].

jump to the corresponding matching positions within the audio recordings from the search results.

4.2 LyricSynchronizer

Fujihara *et al.* developed a music playback interface called LyricSynchronizer based on their algorithm for lyrics synchronization [3]. This music playback interface offers the following two functions: displaying synchronized lyrics, and jump-by-clicking-the-lyrics functions. The former function displays the current position of the lyrics as shown in Figure 9. Although this function resembles the lyrics display in Karaoke, manually labeled temporal information is required in it. Using the latter function, users can change the current playback position by clicking a phrase in the lyrics. This function is useful when users want to listen only to sections of interest.

4.3 SongPrompter

Mauch *et al.* developed a software system called SongPrompter [11] by utilizing their lyrics-and-chord-to-audio alignment method [12]. This system acts as a performance guide by showing lyrics, chords, beats and bar marks along with music playback. Unlike the previous two examples, this software is designed for music performer. Figure 10 shows a screenshot of the system. As can be seen in the figure, both lyrics and chords are shown in the horizontal scrolling bar so that players can play music and sing without memorizing lyrics and chords or turning pages.

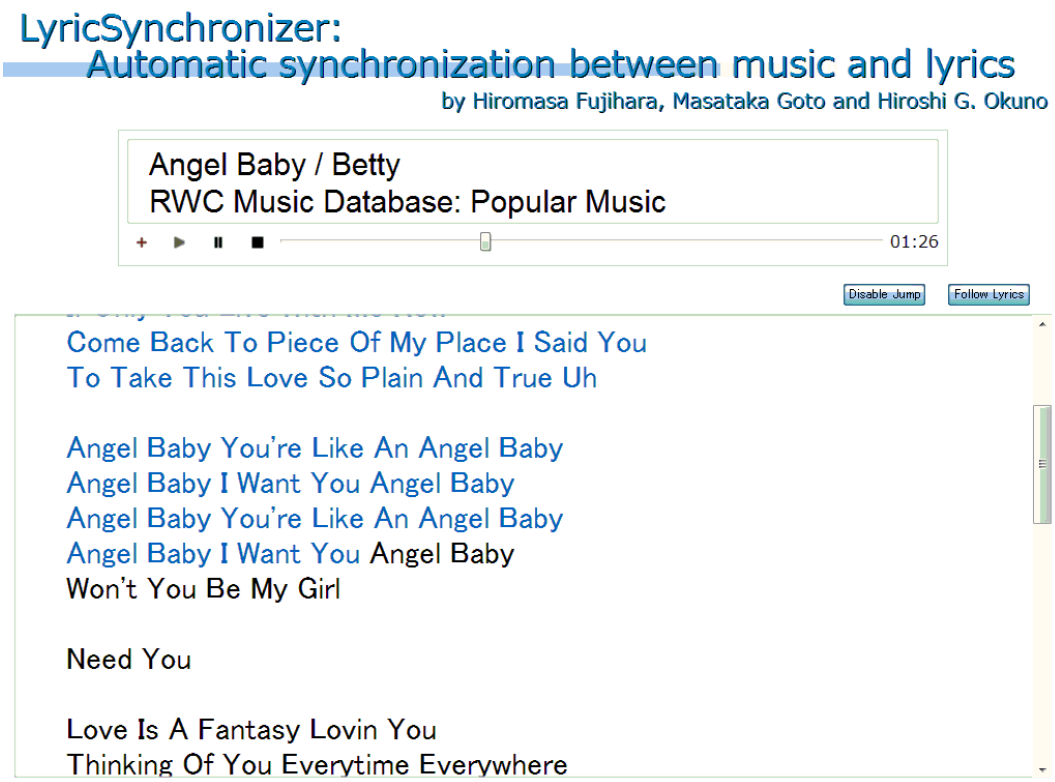


Figure 9 Screenshot of LyricSynchronizer [3].

5 Conclusions

In this paper, we described recent developments in lyrics-to-audio alignment research. We first defined the problem of lyrics-to-audio alignment and then gave an overview of current work. Finally, we introduced several applications of lyrics-to-audio alignment techniques.

Thanks to the advancements of this research fields, it is possible to align lyrics and audio with satisfactory accuracy for songs in which vocals pronounce words clearly and the sounds of vocals are mixed louder. On the other hand, there are still songs of which it is not easy to estimate the correct alignments. As mentioned in Section 3.1, most of lyrics-to-audio alignment techniques have sought to improve their performance by integrating various signal processing and music understanding techniques. This is because singing voices are highly correlated with other elements in music (e.g. melody F0s and chords) and, thus, the understandings of such elements can help aligning lyrics and singing voices.

To advance this field further, we think that the following three approaches can be conceivable:

Integrating of other signal processing and music understanding techniques. We believe that it is a promising direction to integrate fruits of a broader array of research field. For example, recent developments of source separation research can contribute to lyrics-to-audio alignment research. It is also possible to incorporate music classification methods such as genre detection and singer identification to select a model which is most suited for an input song.

A more sophisticated way of integrating information. As a way of integrating various in-

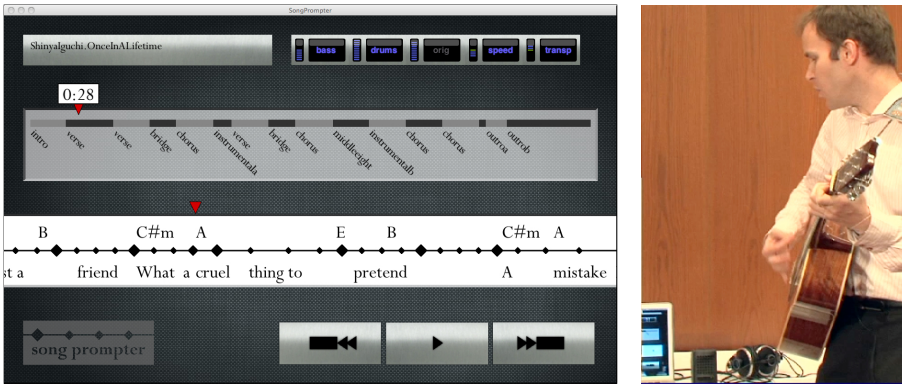


Figure 10 *SongPrompter* interface screenshot and usage example [11].

formation extracted by several music understanding techniques, most of the current studies took a straightforward approach: each music understanding technique was regarded as independent and only the results from different techniques were integrated. However, we believe it is possible to boost the performance by integrating the process of each music understanding technique so that each technique works in a mutually complementary manner.

Practically-oriented approach by utilizing external information. Finally, it is also interesting to incorporate external information available on the Web or other places. This approach, which narrows the range of applicable songs but can lead to interesting applications, was already taken by Müller *et al.* (lyrics-aligned MIDI) [15] and Mauch *et al.* (lyrics with chord annotation) [12] and resulted in the appealing applications as described in the previous section. We think that there could be other sources of information that are easy to obtain and can be beneficial to improve the performance.

References

1 Adam L. Berenzweig and Daniel P. W. Ellis. Locating singing voice segments within music signals. In *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

2 Hiromasa Fujihara, Masataka Goto, Tetsuro Kitahara, and Hiroshi G. Okuno. A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity based music information retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 18:638–648, 2010.

3 Hiromasa Fujihara, Masataka Goto, Jun Ogata, and Hiroshi G. Okuno. LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing*, 5:1252–1261, 2011.

4 Masataka Goto. A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329, 2004.

5 Matthias Gruhne, Konstantin Schmidt, and Christian Dittmar. Phoneme recognition in popular music. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 369–370, 2007.

6 Denny Iskandar, Ye Wang, Min-Yen Kan, and Haizhou Li. Syllabic level automatic synchronization of music signals and text lyrics. In *Proceedings of ACM Multimedia*, pages 659–662, 2006.

7 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.

8 Peter Knees, Markus Schedl, and Gerhard Widmer. Multiple lyrics alignment: Automatic retrieval of song lyrics. In *Proceedings of the 6th International Conference on Music Information Retrieval*, pages 564–569, 2005.

9 Kyogu Lee and Markus Cremer. Segmentation-based lyrics-audio alignment using dynamic programming. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 395–400, 2008.

10 Alex Loscos, Pedro Cano, and Jordi Bonada. Low-delay singing voice alignment to text. In *Proceedings of the International Computer Music Conference 1999 (ICMC99)*, 1999.

11 Matthias Mauch, Hiromasa Fujihara, and Masataka Goto. Song Prompter: An accompaniment system based on the automatic alignment of lyrics and chords to audio. In *Late-breaking session at the 10th International Conference on Music Information Retrieval*, 2010.

12 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), 2012.

13 Annamaria Mesaros and Tuomas Virtanen. Automatic alignment of music audio and lyrics. In *Proceedings of the 11th International Conference on Digital Audio Effects*, pages 1–4, 2008.

14 Annamaria Mesaros and Tuomas Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Advances in Signal Processing*, 2010, 2010.

15 Meinard Müller, Frank Kurth, David Damm, Christian Fremerey, and Michael Clausen. Lyrics-based audio retrieval and multimodal navigation in music collections. In *Proceedings of the 11th European Conference on Digital Libraries (ECDL 2007)*, 2007.

16 Tomoyasu Nakano and Masataka Goto. VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation. In *Proceedings of the 6th Sound and Music Computing Conference*, pages 343–348, 2009.

17 Naoki Nishikawa, Katsutoshi Itoyama, Hiromasa Fujihara, Masataka Goto, Tetsuya Ogata, and Hiroshi G. Okuno. A musical mood trajectory estimation method using lyrics and acoustic features. In *Proceedings of the First International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies (MIRUM 2011)*, pages 51–56, 2011.

18 Tin Lay Nwe and Ye Wang. Automatic detection of vocal segments in popular songs. In *Proceedings of the 5rd International Conference on Music Information Retrieval*, pages 138–145, 2004.

19 Akira Sasou, Masataka Goto, Satoru Hayamizu, and Kazuyo Tanaka. An auto-regressive, non-stationary excited signal parameter estimation method and an evaluation of a singing-voice recognition. In *Proceedings of the 2005 International Conference on Acoustics, Speech, and Signal Processing*, pages I–237–240, 2005.

20 Johan Sundberg. *The Science of Singing Voice*. Northern Illinois University Press, 1987.

21 Motoyuki Suzuki, Toru Hosoya, Akinori Ito, and Shozo Makino. Music information retrieval from a singing voice using lyrics and melody information. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

22 Wei-Ho Tsai and Hsin-Min Wang. Automatic detection and tracking of target singer in multi-singer music recordings. In *Proceedings of the 2004 International Conference on Acoustics, Speech, and Signal Processing*, pages 221–224, 2004.

23 Chong-Kai Wang, Ren-Yuan Lyu, and Yuang-Chin Chiang. An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker. In

Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech2003), pages 1197–1200, 2003.

24 Chi Hang Wong, Wai Man Szeto, and Kin Hong Wong. Automatic lyrics alignment for Cantonese popular music. *Multimedia System*, 4-5(12):307–323, 2007.

25 Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK Book Version 3.4*. Cambridge University Engineering Department, 2006.

Fusion of Multimodal Information in Music Content Analysis*

Slim Essid and Gaël Richard

Institut Télécom, Télécom ParisTech, CNRS-LTCI
37 rue Dareau, 75014 Paris, France
Slim.Essid@telecom-paristech.fr, Gael.Richard@telecom-paristech.fr

Abstract

Music is often processed through its acoustic realization. This is restrictive in the sense that music is clearly a highly multimodal concept where various types of heterogeneous information can be associated to a given piece of music (a musical score, musicians' gestures, lyrics, user-generated metadata, etc.). This has recently led researchers to apprehend music through its various facets, giving rise to *multimodal music analysis* studies. This article gives a synthetic overview of methods that have been successfully employed in multimodal signal analysis. In particular, their use in music content processing is discussed in more details through five case studies that highlight different multimodal integration techniques. The case studies include an example of cross-modal correlation for music video analysis, an audiovisual drum transcription system, a description of the concept of informed source separation, a discussion of multimodal dance-scene analysis, and an example of user-interactive music analysis. In the light of these case studies, some perspectives of multimodality in music processing are finally suggested.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems, I.5 Pattern Recognition

Keywords and phrases Multimodal music processing, music signals indexing and transcription, information fusion, audio, video

Digital Object Identifier 10.4230/DFU.Vol3.11041.37

1 Introduction

While the most natural way to perceive music is through its acoustic rendering, it is clear that it is a highly multimodal concept that can be sensed in a variety of ways: music is materialized in the head of a composer, or a trained musician reading a musical-score; it is translated into sound and motion in a performer's gestures or a dancer's movements and steps; it becomes visual art when it is illustrated by disc cover designs or transformed into an audiovisual production; not to mention its textual dimension that encapsulates not only the lyrics (in sung music) and editorial metadata, but also social web content such as user-tags, reviews, ratings, etc.

Consequently, treating music only through its acoustic realization appears to be quite restrictive, which has led researchers in the general field of music content analysis to apprehend it through its various facets, giving rise to *multimodal music analysis* studies. To our knowledge the earliest contributions along this line dealt with two modalities, that is the audio and score modalities, in order to perform music-to-score matching [10, 66]. The more complex visual modality has not been exploited in music analysis until the late 90s [60],

* This work was partially supported by the European commission with the 3Dlife project

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 37–52

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

in contrast to the speech processing domain where audiovisual speech recognition systems have been imagined in the 80s [59]. Not surprisingly, the earliest works on audiovisual music were dedicated to the analysis of piano music [60], [62], probably due to the possibility to segment the keyboard keys and track the musician's fingers positions on the keyboard more easily than with other instruments.

Since then, our field of interest has seen a variety of multimodal studies spanning a wide range of techniques and applications, an overview of which is proposed in this article. We will first provide a synthetic view of methods that have been successfully employed in multimodal research works in general and discuss their use for music processing. Subsequently, we will discuss a selection of case studies we have contributed to, and highlight the related future research directions that seem promising to us.

2 Multimodal Techniques

Multimodal processing techniques, in general, fall into one of two categories of a binary taxonomy: *early integration techniques* as opposed to *late integration techniques*.[1] The former refers to the process whereby a system directly exploits the "raw" low-level features used to describe each data stream, without any further transformations other than basic postprocessing (typically denoising, normalisation, resampling, etc.). By contrast, the latter is employed to indicate that the joint exploitation of the modalities is performed at a *decision-level*, typically by combining the outputs of intermediate monomodal classifiers. This distinction will be useful to understand the differences between the techniques presented hereafter. Another interesting distinction is the following: the effort of characterizing the "relationships" between the different modalities reflecting the content being analyzed is referred to as *cross-modal processing*, while the problem of efficiently combining the information conveyed by the different modalities (to perform a more thorough analysis of the content) is called *multimodal fusion*. Below we further describe the previous paradigms and discuss their exploitation in the field of music processing.

2.1 Cross-Modal Processing

The relationships between the modalities considered can be expressed in several different ways.

In the first place, when dealing with modalities having a temporal dimension (typically audio signals, video signals, or musical scores), it might be required to temporally align the different data streams in case they are not initially synchronized. In fact, achieving this synchronization may be one's ultimate goal: for instance when dealing with the audio and score modalities, this task is often referred to as *music-to-score alignment* (or *music-to-score synchronization*) [50, 41]. Since the latter is already well covered in other articles of this book, we will here assume that the data streams considered are temporally aligned.

Assuming synchronized features, many proposals have been made to measure a form of dependency between two heterogeneous data streams, part of which remain under-exploited in the music information retrieval community, despite their potential. For the sake of clarity, we make the assumption (without loss of generality) that two streams of data are considered: an audio stream and a video stream. Though the methods presented in the following have

[1] It is worth mentioning that hybrid approaches exist too.

been mainly applied to those two particular modalities, they can be used with any other parallel data streams whose dependency is to be characterized.

A number of techniques have been suggested to map the observed audio and visual feature vectors to a low dimensional space where a *measure of "dependency"* between them can be computed. Let us assume the n observed audio feature vectors $x_a \in \mathbb{R}^{D_a}$ are assembled column-wise in a $(n \times D_a)$-matrix X_a, and the corresponding visual feature vectors[2] $x_v \in \mathbb{R}^{D_v}$ are assembled column-wise in a $(n \times D_v)$-matrix X_v. The methods we describe here aim to find two mappings f_a and f_v (that reduce the dimensions of the audio and visual feature vectors), such that a dependency measure $S_{av}(f_a(X_a), f_v(X_v))$ is maximized. Various approaches can be described using this same formalism. Darrel *et. al.* choose the mutual information [8] as a dependency measure and seek single-layer perceptrons f_a and f_v projecting the audiovisual feature vectors to a 2-dimensional space. Other more popular approaches, for which closed-form solutions can be found, use linear mappings to project the feature streams:

- Canonical Correlation Analysis (CCA), first introduced by Hotelling [33], aims at finding pairs of unit-norm vectors t_a and t_v such that

$$(t_a, t_v) = \arg \max_{(t_a,t_v)\in\mathbb{R}^{D_a}\times\mathbb{R}^{D_v}} \operatorname{corr}(t_a^t X_a, t_v^t X_v)\,. \tag{1}$$

- An alternative to the previous (expected to be more robust than CCA) is Co-Inertia Analysis (CoIA). It consists in maximizing the covariance between the projected audio and visual features:

$$(t_a, t_v) = \arg \max_{(t_a,t_v)\in\mathbb{R}^{D_a}\times\mathbb{R}^{D_v}} \operatorname{cov}(t_a^t X_a, t_v^t X_v)\,. \tag{2}$$

- Yet another configuration known as Cross-modal Factor Analysis (CFA), and found to be more robust than CCA in [45], seeks two matrices T_a and T_v, such that

$$(T_a, T_v) = \arg \max_{(T_a,T_v)} (1 - ||T_a X_a - T_v X_v||_F^2) = \arg \min_{(T_a,T_v)} ||T_a X_a - T_v X_v||_F^2; \tag{3}$$

  with $T_a T_a^t = I$ and $T_v T_v^t = I$. $||X||_F$ denotes the Frobenius norm of matrix X.

Note that the previous techniques can be kernelized to study non-linear coupling between the modalities considered (see for instance [44, 31]).

The interested reader is referred to [33, 31, 45] for further details on these techniques, and to [25] for a comparative study. Examples of applications in the field of music content processing are mentioned in Section 2.4.

2.2 Feature-Level Fusion

Feature-level fusion is the (early integration) process of combining different types of features from different modalities into a common feature representation.

The most basic audiovisual feature fusion approach consists in concatenating the audio and visual feature vectors, x_a and x_v, to form a global feature vector $x_{av} = [x_a, x_v]$. However,

[2] The underlying assumption is that the (synchronized) audio and visual features are extracted at the same rate, which is often obtained by downsampling the audio features or upsampling the video features, or by using temporal integration techniques [40].

the dimensionality of the resulting representation is often too high, leading researchers to resort to dimensionality reduction methods.

A common approach is to use *feature transformation* techniques such as Principal Component Analysis (PCA) [5], Independent Component Analysis (ICA) [63], or Linear Discriminant Analysis [5]. An interesting alternative, is *feature selection* [30] which aims to select only useful descriptors for a given task and discard the others. Indeed, when applied to the feature vectors x_{av}, feature selection can be considered as a feature fusion technique whereby the output will hopefully retain the "best of x_a and x_v" *i.e.* a subset of the most relevant audiovisual features (with respect to the selection criterion).

Nevertheless, the two previous approaches can be considered as limited owing to the different physical nature of the audio-visual features to be combined. In particular, the features do not necessarily live in the same metric space, and are not necessarily extracted from the same temporal segments. Consequently, there has been a number of proposals attempting to address these limitations. One possible approach consists in building separate kernels for different features, before determining new optimal kernels (as convex combinations of the individual ones) in order to use them for classification [70]. Another possible approach of note is the construction of joint audiovisual representations, envisaged as *audiovisual atoms* in [38], and *audiovisual grouplets* in [39], both exploiting audiovisual correlations. The joint audiovisual representation may in particular be built using one of the audiovisual subspace methods described in Section 2.1 (see [45] for an example).

2.3 Decision-level fusion

Late fusion or the idea of combining intermediate monomodal decisions[3] in order to achieve a more accurate multimodal characterization of a content has been explored extensively, under various configurations.

Numerous works rely on *majority voting* procedures whereby final global decisions are made based on a weighted sum of individual voters, each typically corresponding to a decision taken on a particular modality. The weights are often chosen using either heuristics or trial-and-error procedures (see for example [46]). This idea can be better formalized using a Bayesian framework, which allows for taking into account the uncertainty about each classifier's decisions, as done in [36]. Also, solutions to deal with the potential imprecision of some modalities have been proposed using the *Dempster-Shafer* theory [19]. Another widely used strategy consists in using the monomodal classifiers outputs as features, on the basis of which a new classifier, that is expected to optimally perform the desired multimodal fusion, is learned [68].

The previous approaches do not account for the dynamic properties of the media streams considered, nor do they allow for encoding prior knowledge about the dependency structure in the data, in particular the temporal and/or cross-modal dependencies. To this end, sophisticated dynamic classifiers have been utilized, ranging from variants of (multi-stream) Hidden Markov Models (HMM) [28, 52, 43, 1], through more general Dynamic Bayesian Networks (DBN) [6, 27], to even more general graphical models such as Conditional Random Fields (CRF) [41, 3].

[3] These decisions are generally output by previously trained classifiers.

Table 1 Case studies presented. The "Modalities" are the ones taken into account in the corresponding case study; "Cross-modal" indicates whether the method presented performs cross-modal analysis; "Fusion" indicates whether it exploits multimodal fusion and "Section" is where in this chapter the case study is presented.

Case studies	Modalities	Cross-modal	Fusion	Section
Audiovisual correlation in music videos	audio, video	•	○	3.1.1
Audiovisual drum transcription	audio, video	○	•	3.1.2
Music in motion: analyzing dance scenes	audio, motion, depth, video, choreographies	•	•	3.2
Interactive music analysis	audio, human	•	•	3.3
Informed source separation	audio, score, human	•	○	3.4

2.4 Discussion

Many of the techniques mentioned above have been exploited in multimodal music content analysis research. Cross-modal analysis seems to be particularly popular within this domain. For instance, CCA has been used both for studying correlations between sounds and human motion or gestures [54, 55], and correlations between music and words, in view of creating a musically meaningful vocabulary [65]. Also, heuristic rules for the association of higher-level descriptors extracted from different modalities have been employed [4, 21]. In fact, it seems that approaches relying on heuristic rules are mainstream, be it for specific content analysis tasks, such as music video summarization [71], or more general classification problems (see for example [46] where the output of audio and visual classifiers are heuristically combined).

We believe there is a great potential in exploiting the more sophisticated cross-modal techniques and dynamic statistical models previously mentioned to be able to better express one's prior knowledge on the data structure (features dependency, temporal synchronisation, multi-scale effects, higher-level cross-modal concept relationships, etc.) and fully exploit the valuable information that is encoded in it. This of course entails a formalisation effort which is expected to be rewarding both in terms of performance and generalization insofar as the purpose of using common architectures for different applications can be pursued. Modeling the ambiguity and imprecision of intermediate (mono-modal) decisions thanks to the Dempster-Shafer theory of evidence is another interesting idea that is believed to hold much promise.

3 Case Studies

We now present particular multimodal music applications that we have treated in the past few years illustrating the techniques introduced in Section 2. Table 1 gives an overview of these case studies indicating the modalities considered and the class of techniques employed.

3.1 Audiovisual Music

3.1.1 Audiovisual Correlation in Music Videos

The first case study is dedicated to a specific aspect of multimodal signal analysis and aims at exploiting the correlation between the audio and visual modalities in music videos [21].

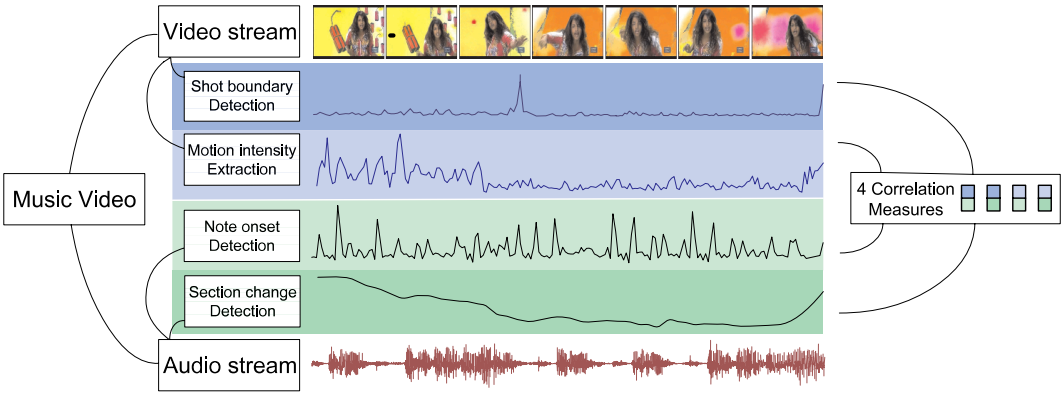


Figure 1 Overview of the audio-visual content structuring system (from [21]).

In the case of music videos, a large palette of semantic relationships between the audio and video streams may be used by the artists at the production stage. For example, mainstream music videos show dancers or performers, but some videos have a narrative content based on higher-level features of the song (such as structure or mood) while others explore new forms of visual metaphors [26, 42, 53].

In this case study (further described in [21]) high-level structures of the audio and video streams are separately extracted in order to measure the correlations between these structures. The objective in such an approach is to characterize the synchrony of significant events and changes in the music and the accompanying images.

It is clear that a large number of salient events can be defined both for audio and visual streams. In music signals, note or chord changes are obviously important events. Thus, an efficient mid-level temporal structuring of a music piece can be achieved by detecting the onsets of such events which coarsely capture the rhythmic properties of the music (many onset detection methods exist and the interested reader may consult the tutorial given in [2]).

In parallel, the events of interest to be extracted from the video include rapid movements such as dance steps, movements of musicians or any action sequence (similarly many approaches exist and such events can be for example detected using motion activity detectors [37]).

At a higher level, a music piece can be temporally segmented in sections, characterized by distinct dynamic, tonal or timbral properties and corresponding to the musical structure of the piece, *i.e.* choruses, verses, fill-ins, etc. Such segments can be either obtained by identifying large blocks in a self-similarity matrix computed on the signal (see for example [58, 7] in the framework of automatic summarization) or by exploiting novelty detection methods which allow for determining boundaries between homogeneous temporal segments [21].

For the video part, the higher level description is obtained by means of a segmentation into shots. In fact, shot changes events are semantically important in the sense that they may be correlated with the rhythm or section changes in the music.

These four segmentation processes produce detection functions (represented in Figure 1) ideally exhibiting peaks whenever an event or section change is detected. The detection functions can be thresholded to obtain the temporal location of salient events and segment boundaries, or directly considered to measure correlations.

The experiments reported in [21] have shown that the correlation between *note onset*

(music) and *shot changes (video)* is particularly appropriate for cross-media authoring or cross-media retrieval applications (e.g. audio retrieval from video or vice-versa video retrieval from audio). In the latter case, it obviously depends on the genre of the music videos. For instance, for narrative videos, where the music video has a strong narrative content and chronology, the proposed mid-level correlations are not adequate since they cannot capture such high level semantic links. Understanding music lyrics, music emotion from audio and video, represent some of the very attractive current and future lines of research in this domain.

This case study is thus an illustration of an exclusively cross-modal application, where multimodal fusion *per se* is not employed, in the sense that one is only interested in detecting the synchrony between the audio and visual streams and not in interpreting or automatically annotating the individual streams. Note that such a matching of the audio and video content at a structural level opens the path for numerous applications, ranging from temporal re-synchronization of mismatched audio and video streams to audio-driven video editing, or soundtrack retrieval by video query.

3.1.2 Audiovisual Drum Transcription

Drum transcription in polyphonic music is a particularly interesting case study for multimodal music analysis. Indeed, for many musical instruments (brass and woods in particular) a small visible movement of the musician's body or fingers may induce a large variation of the produced sound. On the contrary, the nature of the drum kit (e.g. consisting of several drum elements which are physically located at rather different locations) implies that a rather specific movement is needed from the drummer to hit each of the drum elements. It is then expected that multimodality is of great benefit for automatic drum transcription.

Even though a number of studies exist for drum solo transcription (see [18]) or for monomodal (audio-only) drum transcription of polyphonic music signals [24], [57], there has been only a few studies exploiting multimodality. A number of multimodal experiments were conducted by S. Dahl showing the relationship between body movements and emotions in marimba performances or the correlation between video features and musical accents in drumming [9].

In [22], a multimodal system for drum transcription is described exploiting both the video and audio modalities. In this work several early-fusion and late-fusion techniques were evaluated on drum-solos and it was shown that feature-level fusion by simple concatenation of audio and video features can achieve significant improvements compared to either of the monomodal transcription systems. However, with such simple integration schemes, it does not seem obvious that the strength of each modality is well exploited. In this initial system, there is indeed no intent to understand the semantics of the images or to extract higher-level features.

A different strategy is followed in [49] where the video modality is used as a detection process. More precisely (see Figure 2), the video sequence is first analyzed to detect the position of each drum element (drums and cymbals) in the scene, and more specifically the part of the instrument hit by the drum sticks. A geometric criterion is used to detect the drum tops (which are of circular shape). Then, a simple motion intensity feature coupled with foreground object segmentation is used to detect drum strokes on each of the detected drum tops. The transcription is obtained by identifying which drum instrument corresponds to each detected drum top. In parallel, the audio transcription system can also be used, as an additional source of information, to unequivocally assign each detected region to the corresponding drum instrument. Finally, once a video transcription is obtained, it can

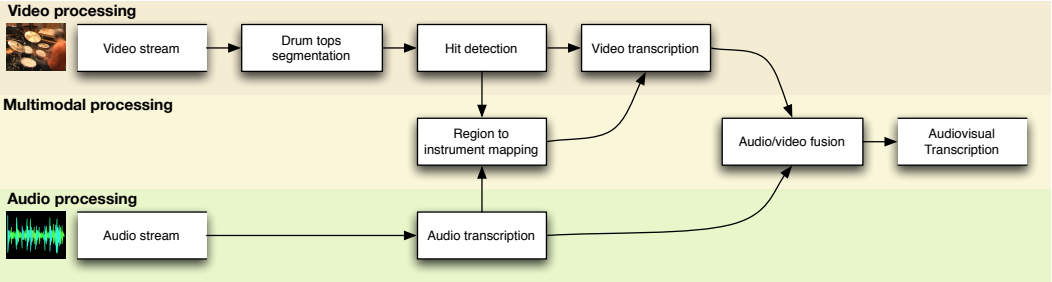


Figure 2 Overview of the audio/video analysis drum transcription system (from [49]).

be fused with an audio transcription or other video transcriptions obtained from different cameras.

This multimodal system outperformed both the monomodal systems and the system based on the traditional early and late fusion methods (the evaluation was performed on the Audiovisual ENST-Drums database [23]). One of the interesting lessons that can be learned from this work is that exploiting high-level information obtained from one modality to drive (or at least help) the processing of the other modality can be a better strategy than merely relying on direct feature-level or decision-level fusion.

3.2 Music in Motion: Analyzing Dance Scenes

Dancing is another manifestation of the multimodal nature of music. Indeed, it can be considered as a form of motion-rendering of music by dancers. For most dance styles, the analysis of a dancer's movements cannot be abstracted from the related music, as the steps and movements of the choreography are expected to be responses to particular musical events, an observation that has been successfully exploited in [61, 11].

We here describe a new multimodal dance dataset that is particularly challenging in terms of open research issues, namely the *3DLife dance dataset*[4] [14].

The dataset consists of multimodal recordings of Salsa dancers, captured at different sites with different pieces of equipment, as illustrated in Figure 3. This includes:

- synchronized 16-channel audio capture of dancers' step sounds, voice and music;
- synchronized 5-camera video capture of the dancers from multiple viewpoints covering whole body, plus 4 non-synchronized additional video captures;
- inertial (accelerometer + gyroscope + magnometer) sensor data captured from multiple sensors on the dancers' bodys;
- depth maps for dancers' performances captured using a Microsoft Kinect;
- original music excerpts;
- different types of ground-truth annotations, for instance, annotations of the music in terms of beats, annotations of the choreographies with step time codes relative to the music and ratings of the dancers' performances (by the Salsa teacher).

Over 20 dancers have been captured, each performing 2 to 5 solo Salsa choreographies among a set of 5 pre-defined ones. The dancers have been instructed to execute these choreographies respecting the same musical timing, *i.e.* all are expected to synchronize

[4] http://perso.telecom-paristech.fr/~essid/3dlife-gc-11/

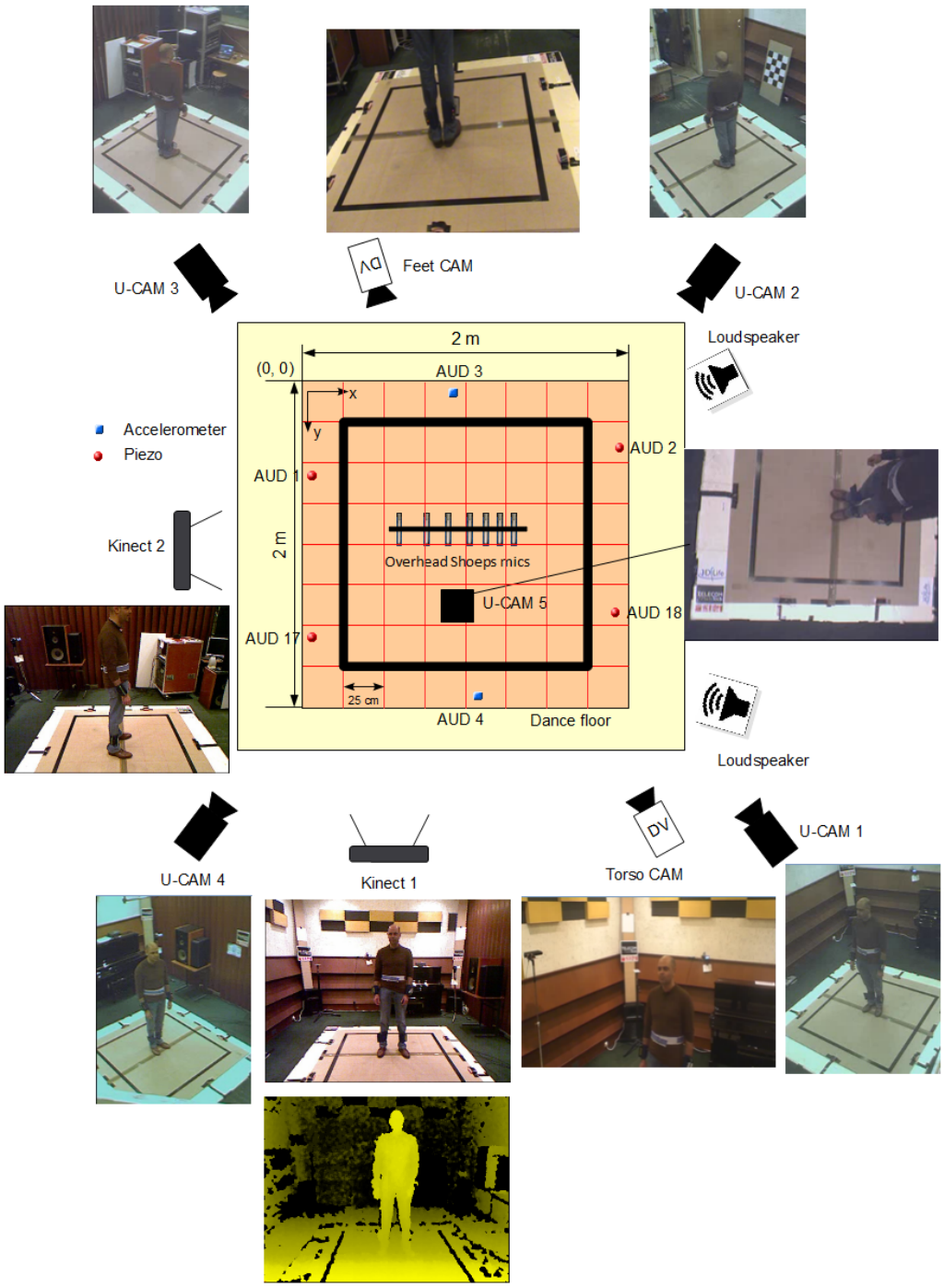


Figure 3 Recording setup at Telecom ParisTech studio.

steps/movements to particular music beats. Salsa music was chosen for this data corpus as it is a music genre that is centered at dance expression, with highly structured, yet not straightforward rhythmic patterns.

The dancers' degree of mastering of Salsa is variable. In particular there are two reference dancers which are considered as the dance teachers whose performances are viewed as the ideal templates to be followed by the other "student-dancers". In fact, this dataset has been designed in view of a broad application scenario that is an online virtual environment for dance teaching (see [14] for more details).

A number of exciting research questions are raised by such a scenario, many of which are intimately connected to multimodal music content analysis issues, in particular:

- multimodal dance performance analysis, including dance step/movement tracking and recognition;
- dance performance rating, which may involve the alignment of a dance-student performance against the teacher's performance for comparison, and/or the analysis of the student's "sense of rhythm" by assessing his/her movements timing with respect to musical timing;
- musical rhythm analysis using the analysis of the timing of a (reliable) dancer's movements;
- automatic dance synthesis for virtual agents.

Some of these tasks have already been approached. For instance encouraging results have been obtained for automatic dance performance rating [13], though more sophisticated approaches are needed towards a more accurate evaluation of a performance that would allow for highlighting a dancer's mistakes across the duration of a choreography.

3.3 User-Interactive Music Analysis

The analysis of some forms of music which cannot be represented by musical scores, in particular *electro-acoustic* music [48], cannot be envisaged without taking into account the viewpoint of a human analyst, for instance a musicologist. This is owing to the highly subjective nature of such an analysis that is linked to high-level cultural and cognitive processes.

Hence, interactive schemes have been considered for the development of electro-acoustic music analysis systems [29]. This scenario is considered as a particularly challenging multimodal scenario in which the music takes two forms: on the one hand, an audio recording (and possibly its visual waveform or spectrogram representation), and on the other hand the analyst's mental perception of the recording. Here the goal is to reach a representation of the recording that matches, insofar as it is feasible, its representation in the mind of the musicologist, in a reasonable period of time. Such a representation often takes a graphical form in which visual objects are chosen by the analyst to represent sound objects. The interested reader is referred to [35] for examples.

In his work, Gulluni has focused on electro-acoustic music pieces that can be represented as the superposition of *sound objects*. Using *relevance-feedback* and *active learning techniques* (see [29] for more details on these techniques), satisfactory performance has been obtained at transcribing such a content into sound objects [29].

Many exciting extensions could be addressed in the continuation of this work. Notably, the user could be equipped with more advanced interfaces, such as EEG[5] headsets, in his/her

[5] ElectroEncephaloGraphy: "the recording of electrical activity along the scalp", see [67] for more details.

interaction with the computer analysis system (which has been so far limited to keyboard and mouse feedback), thus allowing it to take into account their cerebral feedback while listening to the music. Even more general physiological recordings could be employed with the aim to characterize the user's emotional responses to the content, for example ECG, blood pressure, sweat activity, etc.

3.4 Partially-Informed Source Separation

The gradual shift of the general domain of music signal processing from the analysis of isolated notes or monophonic signals to the more challenging and more realistic case of polyphonic music explains the increase of interest for source separation paradigms. Indeed, one of the popular means to deal with polyphony is to first split the signals into individual sources (or components) that can then be individually processed as monophonic signals [51, Section V]. Even if the source separation is, in many situations, not explicit (and may only provide a mid-level representation on which subsequent processing would be easier), it remains a very challenging task for common music recordings (e.g. mono or at best stereo recordings of complex polyphonic signals).

However, performances of source separation systems can be significantly improved by incorporating some prior information about the sources and the mixing process. In unsupervised source separation, this information can be given in form of a specific source model (as for example the source/filter model used in [12] for singing voice separation). But in some cases, one may have access to a richer information that describes the content. This additional information can be provided by a user [64] or by a more or less accurate transcription of the music signal (see for example [32], [16] for score-informed transcription systems). In [64], the goal is to separate the singing voice from the polyphonic recording using some information provided by a user. To that aim, the user mimics the desired source by simply singing or humming the main melody. The source separation is then performed using both the original polyphonic music signal and the user provided input. Since the user's signal is simpler to process (no polyphony) and carries many audible similarities with the original signal in both frequency and temporal behaviors, it greatly helps the source separation.

In some cases, one may have access to a more or less accurate transcription of the polyphonic music by means for example of a MIDI score. The usefulness of this MIDI score (possibly obtained on the Web) depends on its quality or in other words on its accuracy to represent the original recording content.

In real case scenarios, it is usually important to first align the score to the audio recordings (see for example [41, 17, 34, 50]). Then, once aligned, the score is used to guide the source separation. For example, the score is used in [69] for obtaining improved spatial information about the sources in a stereo source separation problem. In other works, the aligned MIDI score is used as priors in the probabilistic model (such as Probabilistic Latent Component Analysis in [20] and [32]). The MIDI score can also be used to define harmonic filters which are built from the fundamental frequencies of each active notes [15]. It is also possible using score informed source separation to focus on specific parts of the music. For example in [16], an automated approach is proposed for the decomposition of a monaural piano recording into sound sources corresponding to the left and the right hands.

The different examples discussed above all exploit another source of information, other than the original audio signal. In all cases, this leads to significant improvements in separation quality.

However, it seems reasonable to assume that the strategy followed in these studies can be extrapolated to a much wider set of information sources including for example the lyrics

of the song, the gender of the singer or possibly his or her emotional state. The availability of cover versions, of some of the separated sources as in recent informed source separation methods ([47],[56]) or of user tags for appropriate source models selection also appear to be extremely valuable sources of information.

4 Conclusion

Signal processing for music analysis is a vibrant and rapidly evolving field of research. The richness and complexity of the music content call for methods that take into account music-specific characteristics including concepts such as pitch, harmony, rhythm, and instrumentation. Nevertheless, a growing trend in music analysis is to tackle the problem in a more global manner and to exploit, whenever possible, the multimodal or multi-faceted aspects of music. In this paper, we have proposed a short synthetic view of some methods that have been successfully used in multimodal signal processing. We have also briefly discussed five case studies as recent examples of successful exploitation of multimodality in music processing. In the light of these case studies, it seems clear that multimodality in music processing is very promising. Although many important challenges in this field are ahead of us, we would like to highlight three main directions for future work:

- **Towards extended multimodality**: Most current studies focus on a limited number of modalities (audio and video, audio and score, audio and tags, ...). Since music is by nature truly multidimensional there is a great interest to incorporate multiple information sources or modality for music analysis tasks (including source separation), such as for example song lyrics, singer/performer's motion and emotional state, user tags, physiological signals (EEG[6], ECG[7], ...), etc.
- **Towards extended cross-modality**: There are no particular reasons why cross-modality should be expressed through simple linear couplings. There is thus a clear perspective to extend the current approaches to non-linear coupling between modalities using for example "kernelized correlations".
- **Towards extended user interaction**: In most studies, the user is not directly involved in the music analysis stage. It seems however important to strengthen the involvement of users by further developing the concept of relevance feedback or active learning which should allow for designing better human-aware multimodal music systems.

5 Acknowledgments

This article is largely based on the works of several students mostly from Telecom ParisTech. We warmly thank Olivier Gillet, Sébastien Gulluni, Kevin Mc Guinness, Romain Hennequin, Cyril Joder and Antoine Liutkus. Part of this work was also conducted with the support from the European Commission with the 3Dlife Network of Excellence [8].

References

1 E. Argones Rua, H. Bredin, C. Garcia Mateo, G. Chollet, and D. Gonzalez Jimenez. Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden markov models. *Pattern Analysis and Applications*, 12(3):271–284, May 2008.

[6] ElectroEncephaloGram
[7] ElectroCardioGram
[8] http://www.3dlife-noe.eu/3DLife/

2 J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, 2005.

3 K. Bousmalis and L. Morency. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 746–752, March 2011.

4 C. Chen, M. Weng, S. Jeng, and Y. Chuang. Emotion-based music visualization using photos. *Advances in Multimedia Modeling*, pages 358–368, 2008.

5 C. Chibelushi, J. Mason, and N. Deravi. Integrated person identification using voice and facial features. *IEE Colloquium on Image Processing for Security Applications (Digest No.: 1997/074*.

6 T. Choudhury, J. Rehg, V. Pavlovic, and A. Pentland. Boosting and structure learning in dynamic bayesian networks for audio-visual speaker detection. In *16th IEEE International Conference on Pattern Recognition*, volume 3, pages 789–794, 2002.

7 M. Cooper and J. Foote. Automatic music summarization via similarity analysis. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

8 T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2006.

9 S. Dahl. The playing of an accent - preliminary observations from temporal and kinematic ana. of percussionists. In *Journal of New Music Research*, volume 29(3), pages 225–234, 2000.

10 R. B. Dannenberg. An on-line algorithm for real-time accompaniment. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 193–198, 1984.

11 Y. Demir, E. Erzin, and Y. Yemez. Evaluation of audio features for audio-visual analysis of dance figures. In *European Signal Processing Conference (EUSIPCO)*, 2008.

12 J.-L. Durrieu, B. David, and G. Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal of Selected Topics in Signal Processing*, October 2011.

13 S. Essid, D. Alexiadis, R. Tournemenne, M. Gowing, P. Kelly, D. Monhagan, P. Daras, A. Dremeau, and N. O'Connor. An advanced virtual dance performance evaluator. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2012.

14 S. Essid, X. Lin, M. Gowing, G. Kordelas, A. Aksay, P. Kelly, T. Fillon, Q. Zhang, A. Dielmann, V. Kitanovski, R. Tournemenne, N. E. O'Connor, P. Daras, and G. Richard. A multimodal dance corpus for research into real-time interaction between humans in online virtual environments. In *ICMI Workshop On Multimodal Corpora For Machine Learning*, Alicante, Spain, November 2011.

15 M. Every and J. Szymanski. A spectral-filtering approach to music signal separation. In *International Conference on Digital Audio Effects, DAFx'04*, Napoli, Italy, October 2004.

16 S. Ewert and M. Müller. Score-informed voice separation for piano recordings. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, Miami, USA, 2011.

17 S. Ewert, M. Müller, and R. B. Dannenberg. Towards reliable partial music alignments using multiple synchronization strategies. In *Proceedings of the International Workshop on Adaptive Multimedia Retrieval (AMR)*, Madrid, Spain, September 2009.

18 D. FitzGerald and J. Paulus. Unpitched percussion transcription. *Signal Processing Methods for Music Transcription*, 2006.

19 S. Foucher, F. Lalibert, G. Boulianne, and L. Gagnon. A dempster-shafer based fusion approach for audio-visual speech recognition with application to large vocabulary french speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

20 J. Ganseman, P. Scheunders, G. Mysore, and J. Abel. Evaluation of a score-informed source separation system. In *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, Utrecht, August 2010.

21 O. Gillet, S. Essid, and G. Richard. On the correlation of automatic audio and visual segmentations of music videos. *IEEE Trans. on Circuit and Systems for Video Technology*, March 2007.

22 O. Gillet and G. Richard. Automatic transcription of drum sequences using audiovisual features. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005.

23 O. Gillet and G. Richard. Enst-drums: an extensive audio-visual database for drum signals processing. *Proceedings of the International Society for Music Information Retrieval Conference*, 2006.

24 O. Gillet and G. Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech and Language Processing*, 16(3):529–540, March 2008.

25 R. Goecke and J. Millar. Statistical analysis of the relationship between audio and video speech parameters for australian english. In *ISCA Tutorial and Research Workshop on Auditory-Visual Speech Processing AVSP 2003*, pages 133–138, September 2003.

26 M. Gondry. *The Work of Director Michel Gondry*. Director's Series, Vol. 3, DVD, Palm Pictures, 2003.

27 J. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes. Dbn based multi-stream models for audio-visual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004.

28 G. Gravier, G. Potamianos, and C. Neti. Asynchrony modeling for audio-visual speech recognition. In *Proceedings of the second international conference on Human Language Technology Research*, pages 1–6, San Diego, California, 2002. Morgan Kaufmann Publishers Inc.

29 S. Gulluni, O. Buisson, S. Essid, and G. Richard. An interactive system for electro-acoustic music analysis. In *International Conference of Music Information Retrieval*, 2011.

30 I. Guyon and A. Elisseeff. An introduction to feature and variable selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

31 D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

32 R. Hennequin, B. David, and R. Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011.

33 H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3 - 4):321 – 377, 1936.

34 N. Hu, R. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, US, October 2003.

35 INA - GRM. Portraits polychromes. `http://www.inagrm.com/accueil/collections/portraits-polychromes`.

36 Y. Ivanov, T. Serre, and J. Bouvrie. Error weighted classifier combination for multi-modal human identification. Technical Report MIT-CSAIL-TR-2005-081, MIT, 2005.

37 S. Jeannin and A. Divakaran. Mpeg-7 visual motion descriptors. *IEEE Trans. Circuits and Systems for Video Technology*, 11:720–724, 2001.

38 W. Jiang, C. Cotton, S. Chang, D. Ellis, and A. Loui. Short-term audiovisual atoms for generic video concept classification. In *Proceedings of the 17th ACM International Conference on Multimedia*, pages 5–14. ACM, 2009.

39 W. Jiang and A. Loui. Audio-visual grouplet: temporal audio-visual interactions for general video concept classification. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 123–132, Scottsdale, USA, 2011.

40 C. Joder, S. Essid, and G. Richard. Temporal integration for audio classification with application to musical instrument classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):174–186, 2009.

41 C. Joder, S. Essid, and G. Richard. A conditional random field framework for robust and scalable audio-to-score matching. *IEEE Transaction on Audio, Speech and Language Processing*, 19(8):2385–2397, November 2011.

42 S. Jonze. *The Work of Director Spike Jonze*. Director's Series, Vol. 1, DVD, Palm Pictures, 2003.

43 E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot. HMM based structuring of tennis videos using visual and audio cues. In *Proceedings of the 2003 International Conference on Multimedia and Expo*, pages 309–312, 2003.

44 P. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(5):365–378, 2000.

45 D. Li, N. Dimitrova, M. Li, and I. Sethi. Multimedia content processing through cross-modal association. In *ACM International Conference on Multimedia*, Berkeley, CA, USA, November 2003.

46 A. Lim, K. Nakamura, K. Nakadai, T. Ogata, and H. Okuno. Audio-visual musical instrument recognition. In *National Convention of Audio-Visual Information Processing Society*, March 2011.

47 A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard. Informed source separation through spectrogram coding and data embedding. *Signal Processing*, September 2011.

48 P. Manning. *Electronic and computer music*. Oxford University Press, Jan 2004.

49 K. McGuinness, O. Gillet, N. O'Connor, and G. Richard. Visual analysis for drum sequence transcription. In *European Signal Processing Conference (Eusipco)*, Poznan, Pologne, sep 2007.

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

51 M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard. Signal processing for music analysis. *IEEE Journal on Selected Topics in Signal Processing*, 5(6):1088–1110, 2011.

52 A. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy. A coupled HMM for audiovisual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2. IEEE, 2002.

53 V. Niches. *Extraordinary Music Videos*. DVD, EAF Music„ 2002.

54 K. Nymoen, B. Caramiaux, M. Kozak, and J. Torresen. Analyzing sound tracings: a multimodal approach to music information retrieval. In *First International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies (MIRUM)*, 2011.

55 H. Ohkushi, T. Ogawa, and M. Haseyama. Music recommendation according to human motion based on kernel CCA-based relationship. *EURASIP Journal on Advances in Signal Processing*, 2011(1):121, December 2011.

56 M. Parvaix and L. Girin. Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1721 –1733, August 2011.

57 J. Paulus and A. Klapuri. Drum sound detection in polyphonic music with hidden markov models. *EURASIP J. Audio Speech Music Process.*, 2009:14:1–14:9, January 2009.

58 G. Peeters, A. L. Burthe, and X. Rodet. Toward automatic music audio summary generation from signal analysis. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

59 E. Petajan, B. Bischoff, D. Bodoff, and N. Brooke. An improved automatic lipreading system to enhance speech recognition. In *CHI '88 Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 19 – 25, 1988.

60 J. Saitoh, A. Kodate, and H. Tominaga. Integrated data processing between image and audio - musical instrument (piano) playing information processing. In *6th International Conference on Image Processing and its Applications*, pages 438–442 vol.1, July 1997.

61 T. Shiratori, A. Nakazawa, and K. Ikeuchi. Detecting dance motion structure through music analysis. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

62 P. Smaragdis and M. Casey. Audio/visual independent components. In *International Symposium on Independent Component Analysis and Blind Source Separation (ICA)*, 2003.

63 P. Smaragdis and C. M. Audio visual independent components. In *International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 709–714, 2003.

64 P. Smaragdis and G. J. Mysore. Separation by "humming": User-guided sound extraction from monophonic mixtures. In *IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, pages 69–72, 2009.

65 D. Torres, D. Turnbull, L. Barrington, and G. Lanckriet. Identifying words that are musically meaningful. In *International Conference of Music Information Retrieval*, pages 405–410, 2007.

66 B. Vercoe. The synthetic performer in the context of live performance. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 199–200, 1984.

67 Wikipedia. Electroencephalography. http://en.wikipedia.org/wiki/Electroencephalography.

68 P. Wilkins, T. Adamek, D. Byrne, G. Jones, H. Lee, G. Keenan, K. Mcguinness, N. E. O'Connor, A. F. Smeaton, A. Amin, Z. Obrenovic, R. Benmokhtar, E. Galmar, B. Huet, S. Essid, R. Landais, F. Vallet, G. T. Papadopoulos, S. Vrochidis, V. Mezaris, I. Kompatsiaris, E. Spyrou, Y. Avrithis, R. Morzinger, P. Schallauer, W. Bailer, T. Piatrik, K. Chandramouli, E. Izquierdo, M. Haller, L. Goldmann, A. Samour, A. Cobet, T. Sikora, and P. Praks. K-space at TRECVid 2007. In *TREC Video Retrieval Evaluation: TRECVID 2007*, November 2007.

69 J. Woodruff, B. Pardo, and R. Dannenberg. Remixing stereo music with score-informed source separation. In *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, Victoria, October 2006.

70 Y. Wu, C.-Y. Lin, E. Chang, and J. Smith. Multimodal information fusion for video concept detection. In *International Conference on Image Processing*, pages 2391 – 2394, October 2004.

71 C. Xu, X. Shao, N. Maddage, and M. Kankanhalli. Automatic music video summarization based on audio-visual-text analysis and alignment. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 361–368. ACM, 2005.

A Cross-Version Approach for Harmonic Analysis of Music Recordings

Verena Konz and Meinard Müller*

Saarland University and MPI Informatik
Campus E1-4, 66123 Saarbrücken, Germany
vkonz@mpi-inf.mpg.de, meinard@mpi-inf.mpg.de

Abstract

The automated extraction of chord labels from audio recordings is a central task in music information retrieval. Here, the chord labeling is typically performed on a specific audio version of a piece of music, produced under certain recording conditions, played on specific instruments and characterized by individual styles of the musicians. As a consequence, the obtained chord labeling results are strongly influenced by version-dependent characteristics. In this chapter, we show that analyzing the harmonic properties of several audio versions synchronously stabilizes the chord labeling result in the sense that inconsistencies indicate version-dependent characteristics, whereas consistencies across several versions indicate harmonically stable passages in the piece of music. In particular, we show that consistently labeled passages often correspond to correctly labeled passages. Our experiments show that the cross-version labeling procedure significantly increases the precision of the result while keeping the recall at a relatively high level. Furthermore, we introduce a powerful visualization which reveals the harmonically stable passages on a musical time axis specified in bars. Finally, we demonstrate how this visualization facilitates a better understanding of classification errors and may be used by music experts as a helpful tool for exploring harmonic structures.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems, I.5 Pattern Recognition

Keywords and phrases Harmonic analysis, chord labeling, audio, music, music synchronization, audio alignment

Digital Object Identifier 10.4230/DFU.Vol3.11041.53

1 Introduction

Automated chord labeling, which deals with the computer-based harmonic analysis of audio recordings, is one of the central tasks in the field of music information retrieval (MIR) [2, 3, 4, 6, 8, 11, 13, 14, 19, 20, 22, 23]. Harmony is a fundamental attribute of Western tonal music and the succession of chords over time often forms the basis of a piece of music. Thus, chord progressions constitute a powerful mid-level representation for the underlying musical signal and can be applied for various MIR tasks.

* This work has been supported by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI). Meinard Müller is now with Bonn University, Department of Computer Science III, Germany.

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 53–72

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

The evaluation of chord labeling procedures is typically performed on large audio collections, where the automatically extracted chord labels are compared to manually generated ground truth annotations. Here, a piece to be analyzed is typically represented by an audio recording, which possesses version-dependent characteristics. For example, specific instruments are used, which have instrument-dependent sound properties, e. g., concerning the energy distributions in the harmonics. Similarly, room acoustics and other recording conditions may have a significant impact on the audio signal's spectral properties. Finally, by emphasizing certain voices or suppressing others, a musician can change the sound in order to shape the piece of music. As a consequence, the chord labeling results strongly depend on specific characteristics of the considered audio recording. Another major problem arises from the fact, that audio-based recognition results refer to the physical time axis given in seconds of the considered audio recording, whereas score-based analysis results obtained by music experts typically refer to a musical time axis given in bars. This simple fact alone makes it often difficult to get musicologists involved into the evaluation process of audio-based music analysis. For example, for the evaluation of chord labeling procedures, ground truth annotations are required. While the manual generation of audio-based annotations is a tedious and time-consuming process musicians are trained to derive chord labels by means of printed sheet music. Such labels, however, are only of limited use for the evaluation of audio-based recognition results. First research efforts have been directed towards the use of score-based ground truth labels for audio-based chord recognition, where it turned out that incorporating such ground truth labels may significantly improve machine learning methods for chord recognition [12, 15].

In this chapter, we build upon a cross-version chord recognition approach previously suggested in [10]. By exploiting the fact that for a musical work there often exist a large number of different audio recordings as well as symbolic representations, we analyze the available versions independently using some automated chord labeling procedure and employ a late-fusion approach to merge the version-dependent analysis results. Here, the idea is to overcome the strong dependency of chord labeling results on a specific version. We show that by using such a cross-version approach one can achieve a stabilization of the chord labeling results. In particular, we observe that more or less random decisions in the automated chord labeling typically differ across several versions. Such passages often correspond to harmonically instable passages leading to inconsistencies. In contrast, consistencies across several versions typically indicate harmonically stable passages. As one main contribution, we show that consistently labeled passages often correspond to correct labeling results. Consequently, one can exploit the consistency information to significantly increase the precision of the result while keeping the recall at a relatively high level, which can be regarded as a stabilization of the labeling procedure. Furthermore, we show that our cross-version approach is conceptually different to a constraint-based approach, where only chord labels are considered that are particularly close to a given chord model. Unlike our cross-version approach, using such simple constraints leads to a significant loss in recall. As another contribution, we describe how to transform the time axis of analysis results obtained from audio recordings to a common musical time axis given in bars. This not only facilitates a convenient evaluation by a musicologist, but also allows for comparing analysis results across different recorded performances.

Finally, we introduce a powerful visualization which is based on the cross-version chord labeling (another interesting approach for visualizing harmonic structures of tonal music has been suggested in [21]). The cross-version visualization indicates the harmonically stable

passages in an intuitive and non-technical way leading the user to passages dominated by a certain key also referred to as tonal centers. Furthermore, in the case that score-based ground truth labels are also provided, the visualization allows for an in-depth error analysis of chord labeling procedures, which deepens the understanding not only for the employed chord recognizer but also for the music material. Additionally, we exemplarily show how the cross-version visualization may serve musicologists as a helpful tool for exploring harmonic structures of a piece of music.

The remainder of this chapter is organized as follows. First, in Section 2 we give an overview of the cross-version chord labeling framework. In Section 3 we show that using a cross-version approach a stabilization of the chord labeling results can be achieved. Afterwards, in Section 4 we exemplarily demonstrate how the cross-version visualization may be used as a supportive tool for exploring harmonic structures before concluding in Section 5 with open problems and future work.

2 Cross-Version Framework

In this section, we describe the cross-version chord labeling procedure following a similar approach as introduced in [10]. Figure 1 shows the employed procedure in a schematic overview. At this point, we emphasize that our approach is not meant to be of technical nature, and we refer to [3, 13] for an overview of state-of-the-art chord labeling procedures. Instead, we introduce a simple yet powerful paradigm which exploits the availability of different versions of a given piece of music.

In the following, we first give a short introduction to music synchronization and describe how synchronization procedures can be used to transform the time axis of audio-based analysis results to a performance-independent musical time axis. Afterwards, we present the employed chord labeling procedure before introducing the concept of cross-version chord labeling. Finally, by means of several music examples, we illustrate the usefulness of our cross-version visualization.

2.1 Synchronization

In the context of the presented cross-version chord labeling approach the concept of music synchronization is of particular importance. In general, the goal of music synchronization is to determine for a given region in one version of a piece of music the corresponding region within another version [9, 17]. Most synchronization algorithms rely on some variant of dynamic time warping (DTW) and can be summarized as follows. First, the two given versions of a piece of music are converted into feature sequences, say $X := (x_1, x_2, \ldots, x_N)$ and $Y := (y_1, y_2, \ldots, y_M)$, respectively. In the synchronization context, chroma features[1] have turned out to yield robust mid-level representations even in the presence of significant musical variations [1, 7, 17, 18]. Chroma features show a high degree of invariance towards changes in timbre and instrumentation while closely correlating to the harmonic progression of the piece of music. From the feature sequences, an $N \times M$ cost matrix C is built up

[1] Implementations of various chroma feature variants are available at www.mpi-inf.mpg.de/resources/MIR/chromatoolbox/, see also [18].

by evaluating a local cost measure c for each pair of features, i. e., $C(n, m) = c(x_n, y_m)$ for $n \in [1 : N] := \{1, 2, \ldots, N\}$ and $m \in [1 : M]$. Then, a cost-minimizing alignment path, which constitutes the final synchronization result, is computed from C via dynamic programming. For a detailed account on DTW and music synchronization we refer to [9, 17] and the references therein. Based on this general strategy, we employ a multiscale synchronization algorithm based on high-resolution audio features as described in [5]. This approach, which combines the high temporal accuracy of onset features with the robustness of chroma features, generally yields robust music alignments of high temporal accuracy.

2.2 Musical Time Axis

The alignment techniques can be used to transform the time axis of audio-based analysis results to a common musical time axis, see Figure 1 for an overview. To this end, we assume that for a certain piece of music we are given a MIDI representation of the musical score, where the MIDI time axis follows a musically meaningful time axis in bars. Such a MIDI file can be obtained by automatically exporting a score in computer-readable format, which in turn can be generated by applying OMR (optical music recognition) software to scanned sheet music, see Figure 1a. Now, given an audio recording of the same piece of music, one can apply music synchronization procedures to establish temporal links between the timelines of the MIDI representation and the audio version. This linking information allows for transferring bar or beat positions from the MIDI timeline to corresponding time positions (given in seconds) of the audio timeline. Then, the audio timeline can be partitioned into segments each corresponding to e. g. one musical beat or bar. Based on this musically meaningful segmentation, beat- or bar-synchronous audio features can be determined. Then each feature vector corresponds to a musically meaningful time unit that is independent of the respective recorded performance. We will use such synchronized features to directly compare the chord labeling results across the different versions.

2.3 Chord Labeling

The chord labeling is then performed on the basis of the synchronized chroma features, where we furthermore apply a tuning estimation to balance out possible deviations of the performances from standard tuning [7, 13]. Note that numerous chord labeling procedures have been described in the literature. State-of-the-art chord recognizers typically employ statistical models such as hidden Markov models [11, 22, 23] or more general graphical models [13] to incorporate smoothness priors and temporal continuity into the recognition process. Since the respective chord labeling procedure is not in the focus of this chapter, we use a basic template-based chord labeling procedure [6], which better illustrates the kind of information that is enhanced and stabilized by our cross-version strategy. However, note that more complex chord recognizers can be used instead.

In the following, we consider 24 chord categories comprising the twelve major and the twelve minor chords, following the conventions as used for MIREX 2010 [16]. Let Λ denote the set of these 24 categories, then for each $\lambda \in \Lambda$ we define a binary template $\mathbf{t}_\lambda$ that corresponds to the respective chord. The template-based chord labeling procedure consists in assigning to each frame (here, exemplarily, we use a bar-wise frame level) the chord label that minimizes a predefined distance d (in our implementation, we use the cosine distance)

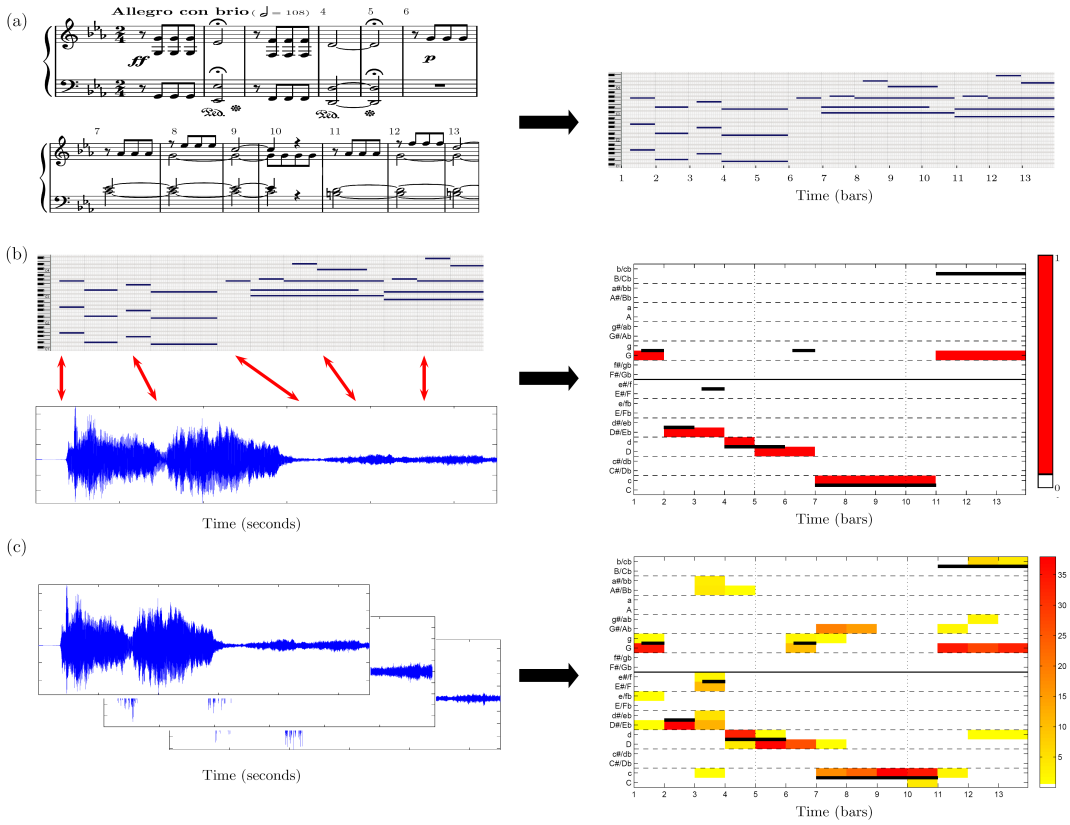


Figure 1 Schematic overview of the employed cross-version framework. Here, the beginning of Beethoven's Fifth (bb. 1-13) is used as an example. (a) Export of the score to a neutral MIDI representation. Here, the score corresponds to a piano reduction of Beethoven's Fifth. (b) Visualization of the automatically derived chord labels for a specific audio recording. The time axis in bars is obtained by synchronizing the audio recording with the MIDI representation. The horizontal black lines in the visualization represent the bassline extracted from the MIDI representation. (c) Cross-version visualization (38 different audio recordings). The horizontal black lines in the visualization represent the bassline extracted from the MIDI representation.

between the corresponding template and a given feature vector referred to as x:

$$\lambda_x := \underset{\lambda \in \Lambda}{\operatorname{argmin}}\ d(\mathbf{t}_\lambda, x). \tag{1}$$

As result, we obtain for each audio version a sequence of automatically extracted bar-wise chord labels. Figure 1b shows the automatically extracted chord labels for a specific audio recording of the first 13 bars of Beethoven's Symphony No. 5, Op. 67, the so-called Beethoven's Fifth. The vertical axis represents the 24 chord categories, where major and minor chords with the same root note are visualized next to each other. Capital letters correspond to major chords, whereas lower case letters correspond to minor chords. The horizontal axis represents the time axis given in bars. The automatically derived chord labels are shown in red, e. g., the chord label for bar 1 corresponds to `G` major, whereas the chord label for bar 2 corresponds to `E♭` major. As the bassline of a harmonic progression plays an important role for the understanding of harmonic structures, we have visualized it as an additional information in the middle of the corresponding major and minor chord having the bassline

as root note. The bassline is automatically extracted from the MIDI representation by determining the lowest of all present MIDI notes at every point in time.

2.4 Cross-Version Chord Labeling

As mentioned in the introduction, the chord labeling results not only depend on the piece of music but also on the acoustic and artistic characteristics of the specific audio recording. To alleviate the dependence on such characteristics, one can exploit the fact that for classical pieces of music usually many different recorded performances exist. Here, our idea is to perform the chord labeling across several versions of a given piece of music and then to resolve the dependency of the chord labels on a specific version by using some kind of late-fusion strategy. Since the automatically extracted chord labels for the different performances are given bar-wise, one can overlay the performance-specific chord labels for all considered recorded performances resulting in a cross-version visualization. Figure 1c shows a cross-version visualization for the beginning of Beethoven's Fifth (bb. 1-13), where 38 different performances are considered. The color-scale ranging from bright yellow to dark red indicates the degree of consistency of the chord labels across the various performances, where red entries point to consistencies and yellow entries to inconsistencies. For example, bar 2 is labeled highly consistently, whereas bar 3 is labeled inconsistently across the considered performances.

In this way, the cross-version visualization directly reveals chord label consistencies and inconsistencies across the different performances giving a deeper insight into the chord labeling procedure as well as the underlying music material. As we will show, consistently labeled passages generally correspond to harmonically stable passages, which are clearly dominated by a certain key. In some cases, consistencies may also point to consistent misclassifications which might be taken as an indicator for inadequacies of the underlying chord labeling model. For example, considering only 24 major and minor chords, it is obvious that more complex chords such as, e. g., diminished chords can not be captured. In contrast, inconsistencies generally point to harmonically instable passages or ambiguities in the underlying music material. For example, incomplete chords as well as additional notes such as trills, appoggiaturas or suspended notes lead to chord ambiguities causing an inconsistent labeling across the different performances.

2.5 Examples

To illustrate our cross-version approach, we now discuss some real-world music examples. We first refer to the introductory bars of Beethoven's Fifth (see Figure 1). Figure 1b shows the visualization of the automatically derived chord labels for a specific audio recording. Following the time axis in bars, the visualization allows for a direct comparison to the score. As the score reveals the first five bars (bb. 1-5) do not contain complete triads. Instead, the characteristic "fate motif" appears, which is presented in octaves in unison. The visualization shows that the automatically derived chord labels for these introductory bars, aside from bar 3, are meaningful in the sense that they represent chords having the presented note of the respective bar as root note. However, in bar 3, where `f` is played in unison, `E♭` major is detected. This might be an indicator for inaccuracies in the synchronization since the previous bar (b. 2) is dominated by the note `e♭`. The same problem appears in bar 6. Bars 7-10 are then labeled as `C` minor. A closer look at the score reveals that in this passage

(bb. 8-10) C minor is clearly present. However, in the beginning of this passage (b. 7) C minor with suspended sixth (a♭) leads into the C minor chord (bb. 8-10). In fact, C minor with suspended sixth corresponds to the notes of A♭ major. However, the suspended sixth (a♭) is played in a very soft way in the considered recording, which might be the reason for the detection of C minor. Bars 11-13 then are labeled in a meaningful way as G major.

The cross-version visualization (Figure 1c) now directly reveals consistently and inconsistently labeled passages. For example, one observes the following highly consistently labeled passages, which may correspond to harmonically stable passages: bars 1-2, 4-5 and 8-13. As previously described, bars 1-2 and 4-5 refer to the fate motif in unison, thus not containing complete triads. These bars are now consistently labeled as a chord having the respective note of the considered bar as root note. Comparing bars 8-13 to the score shows that they indeed correspond to passages being clearly dominated by a certain harmony. Bars 8-10 are consistently labeled correctly as C minor reflecting the harmonic stability of this passage, which is clearly dominated by a C minor triad. Similarly, bars 11-13 are correctly identified by the visualization as harmonically stable, being dominated by G major. In contrast, one directly observes that bar 3 is labeled inconsistently. This inconsistent labeling may be due to local inaccuracies in the underlying synchronization procedure. For a larger amount of recordings this bar is labeled as F major (or as E♭ major) having as root the note presented in unison in this bar (or in the previous bar). In fact, bar 3 was already misclassified as E♭ major considering a single audio recording before. The cross-version visualization now clearly identifies this bar to be problematic in view of the underlying synchronization procedure. Finally, bar 7 attracts attention since it is labeled for approximately half of the recordings as C minor and as A♭ major for the other half. Here, C minor with suspended sixth (a♭) is present, which indeed sounds equivalently to A♭ major. Since the suspended a♭ is usually played in a soft way, for many recordings (including the previously discussed specific recording) this bar is misclassifed as C minor. However, the cross-version visualization shows that for the largest part of recordings this bar is correctly classified (with regard to the sound) as A♭ major.

As the previously discussed example shows, ground truth data is not necessarily needed to derive valuable information from the cross-version visualization concerning the employed chord labeling procedure as well as the underlying music material. However, assuming the case that score-based ground truth labels are provided by a trained musician, this information can be easily incorporated into our cross-version approach, see Figure 2. In this way, errors (deviations from the ground truth) can be subdivided into errors being specific to a certain audio version (inconsistent misclassifications) and errors independent of a specific version (consistent misclassifications). While inconsistent misclassifications may point to ambiguities in the underlying music material, consistent misclassifications may point to inadequacies in the underlying chord labeling framework. In the following, we illustrate such an in-depth error analysis by means of two examples. The score-based ground truth annotations used in our experiments have been generated by a trained musician on the bar-level using the shorthands and conventions proposed by Harte et al. [8].

Figure 2 shows the cross-version visualization for a different excerpt of Beethoven's Fifth (bb. 40-47). On the left, the previously introduced visualization is shown, where the automatically derived cross-version chord labels are visualized without considering ground truth chord labels. On the right, an extension of this cross-version visualization is presented, where the cross-version chord labels are compared to score-based ground truth labels. In this visualization we now distinguish two different color scales: one color scale ranging from dark blue to bright green and the previously introduced color scale ranging from dark red

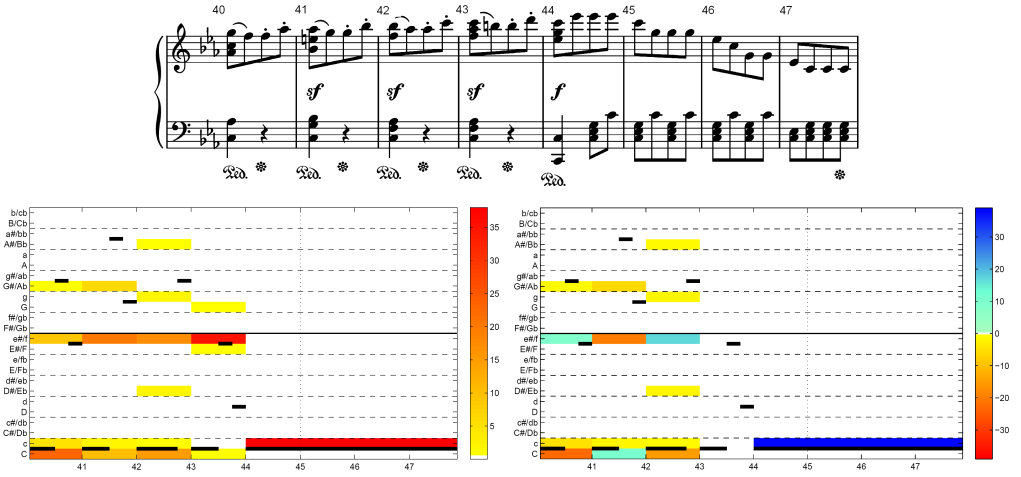


Figure 2 Cross-version visualization for Beethoven's Fifth (bb. 40-47). Here, 38 different audio recordings are considered. **Left:** Cross-version visualization of the automatically derived chord labels. **Right:** Cross-version visualization, where the automatically derived chord labels are overlayed with score-based ground truth chord labels.

to yellow. The first color scale from blue to green serves two purposes. Firstly, it encodes the score-based ground truth chord labels. Secondly, it shows the degree of consistency between the automatically generated audio labels and the score labels. For example, the dark blue entries in bars 44-47 show, that a `C` minor chord is specified in the score-based ground truth labels, and all automatically derived chord labels coincide with the score label here. In contrast, the bright green entry in bar 40 shows that the score-based chord label corresponds to `F` minor, but most of the automatically derived chord labels differ from the score label, specifying a `C` major chord. Analogously, the second color scale from dark red to yellow also fulfills two purposes. Firstly, it encodes the automatically derived chord labels that differ from the score-based labels. Secondly, it measures the universality of an error. For example, in bars 44-47 there are no red or yellow entries, since the score-based labels and the automatically derived labels coincide here. However, in bar 40 most automatically derived chord labels differ from the score-based labels. Here most chord labels specify a `C` major chord.

The cross-version visualization of the automatically derived chord labels (see Figure 2, left) reveals two highly consistently labeled passages: bar 43, labeled highly consistently as `F` minor, and bars 44-47, which are labeled as `C` minor across all considered recorded performances. Comparing to the score, bars 44-47 indeed turn out to be a harmonically stable passage which is clearly dominated by `C` minor. Consequently, this highly consistently labeled passage is labeled correctly, which is shown in the visualization, where the automatically derived chord labels are compared to score-based ground truth labels (see Figure 2, right). In contrast, bar 43 is labeled consistently as `F` minor (see Figure 2, left), but comparing to the score one finds out that besides of an `F` minor chord two additional notes (`b` and `d`) are contained in this bar, suggesting the dominant `G` major. Therefore, a clear assignment of a triad is not possible on the bar level. This is also the reason that there is no score-based label assigned to this bar in the ground truth annotation (see Figure 2, right). The remaining bars are labeled rather inconsistently indicating harmonic instability or ambiguities in the underlying music material (see Figure 2, left). A closer look at the score reveals that these

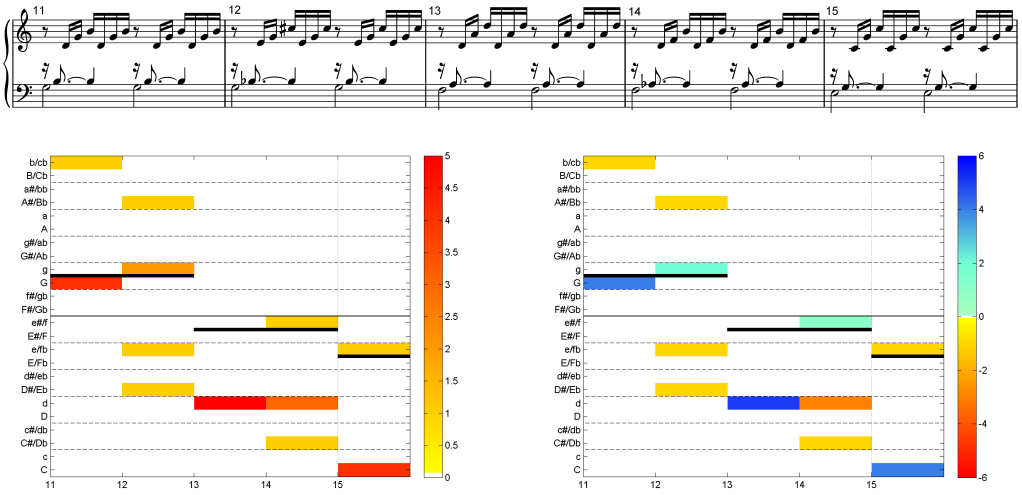


Figure 3 Cross-version visualization for Bach's Prelude BWV 846 in C major (bb. 11-15). Here, five different audio recordings are considered. **Left:** Cross-version visualization of the automatically derived chord labels. **Right:** Cross-version visualization, where the automatically derived chord labels are overlayed with score-based ground truth chord labels.

bars are characterized by suspended notes on the first beat. These additional notes which do not belong to the underlying chords are mainly responsible for the inconsistent labeling. The comparison with the score-based ground truth annotation reveals that for bars 40 and 41 indeed most of the automatically derived chord labels differ from the ground truth annotation (see Figure 2, right).

Figure 3 shows the cross-version visualization for an excerpt of Bach's Prelude BWV 846 in C major (bb. 11-15), where five different recorded performances are considered. The visualization reveals 3 bars which are labeled correctly with high consistency (b. 11, b. 13, and b. 15) and two bars, which are misclassified for most of the considered audio versions (b. 12 and b. 14). Comparing to the score one finds out that the correctly labeled passages indeed correspond to bars, where clear major or minor chords are present. In contrast, bars 12 and 14 are problematic in the sense that they contain diminished seventh chords which can not be assigned in a meaningful way to one of the considered 24 major and minor chords, thus producing misclassifications. In this case, an extension of the considered chord categories to also include diminished seventh chords might solve the problem.

3 Stabilizing Chord Labeling

In this section we show that analyzing the harmonic properties of several audio versions synchronously stabilizes the chord labeling result in the sense that inconsistencies indicate version-dependent characteristics, whereas consistencies across several versions indicate harmonically stable passages in the piece of music. To this end, we introduce a cross-version voting strategy and compare it with a simple constraint-based strategy using a single version. This comparison demonstrates that our voting strategy is conceptually different from simply imposing stricter conditions in the template-based approach. The two strategies are illustrated by means of the first 19 bars of Beethoven's Piano Sonata Op. 27 No. 2, the so-called Moonlight Sonata, see Figure 4.

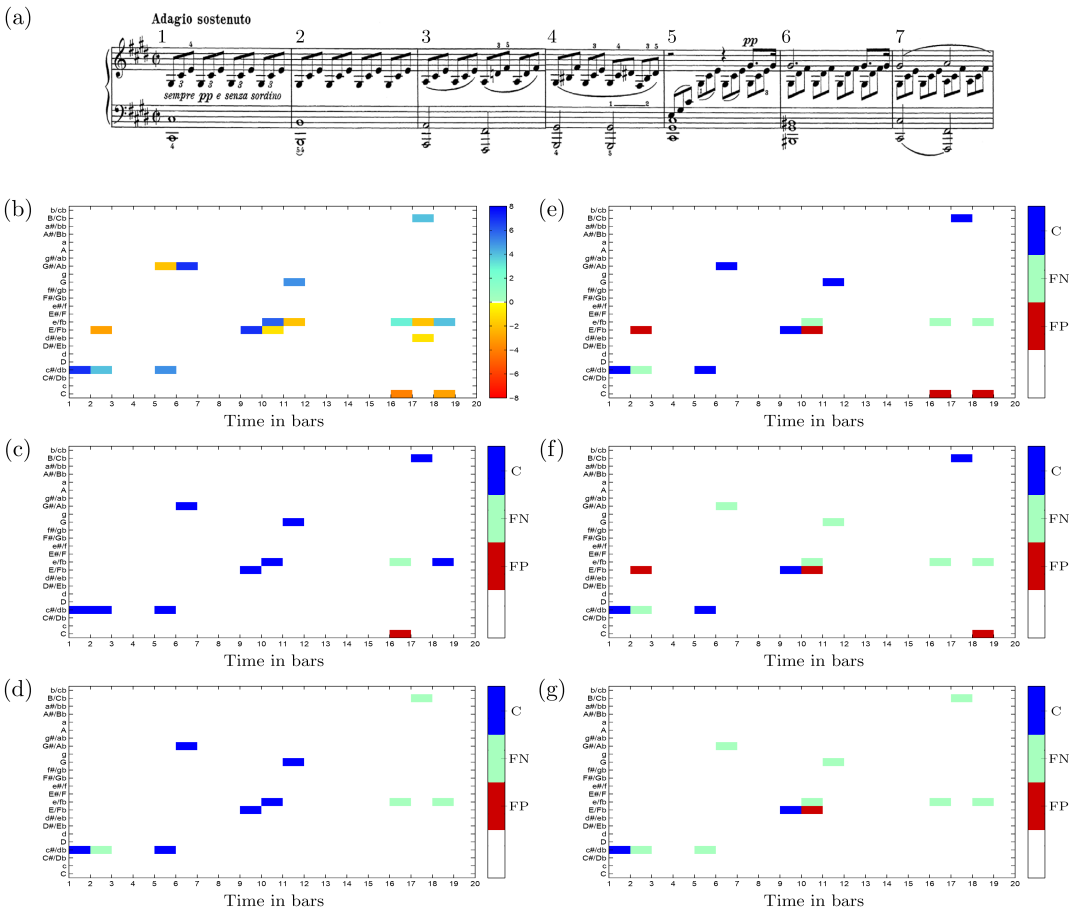


Figure 4 Visualization of the chord labeling result for Beethoven's Moonlight Sonata (bb. 1-19). In the left column (b-d) the cross-version voting strategy is used considering seven performances, whereas in the right column (e-g) the constraint-based strategy is used considering only a single audio recording (Barenboim). Bars, for which no score-based ground truth label exists (since the clear assignment of a harmony is not possible), are left unconsidered in the evaluation. (a) Score of bars 1-7. (b) Visualization of consistencies and inconsistencies in the cross-version analysis. (c) Cross-version majority voting strategy. (d) Cross-version voting strategy with $\nu = 0.5$. (e) Basic strategy. (f) Constraint-based strategy with $\gamma = 0.3$. (g) Constraint-based strategy with $\gamma = 0.1$.

3.1 Cross-Version Voting Strategy

By overlaying the chord labeling results as described in Section 2.4 for the first 19 bars of Beethoven's Moonlight Sonata considering seven different audio versions, we obtain a cross-version visualization, see Figure 4b. The cross-version strategy now reveals consistencies and inconsistencies in the chord labeling across all audio versions. For example, one directly notices that the misclassification in bar 10, when considering a specific audio version (see Figure 4e), seems to be version-dependent. Considering several audio versions, bar 10 is more or less consistently labeled correctly as `E` minor. In contrast, a more consistent misclassification (`C` major instead of `E` minor was labeled for four versions) can be found in bar 16.

In the following experiment, we investigate to which extent the consistency information across several audio versions may be exploited to stabilize chord labeling. In the *majority voting strategy* we keep for each bar exactly one of the automatically extracted chord labels, namely the most consistent chord label across all versions. All remaining audio chord labels are left unconsidered in the evaluation. This results in a visualization which is shown in Figure 4c. Blue entries (correct: C) now indicate areas, where the audio chord label agrees with the ground truth chord label. In contrast, green and red entries encode the differences between the chord labels. Here, red entries (false positives: FP) correspond to the audio chord labels, whereas green entries (false negatives: FN) correspond to the ground truth labels. As one directly notices, besides one misclassification in bar 16, the above mentioned highly consistent error, all chords are now correctly classified resulting in a significant increase of precision.

In the next step, we further constrain the degree of consistency by introducing a consistency parameter $\nu \in [0, 1]$. To this end, we consider only bars which are labeled consistently for more than $(\nu \cdot 100)\%$ of the audio versions. All other bars are left unannotated. For example, $\nu = 0.5$ signifies that we keep in the evaluation only passages, where for more than 50% of the audio versions the extracted chord labels agree. Figure 4d shows the visualization of the chord labeling result for $\nu = 0.5$, where the voting procedure succeeds in eliminating all misclassifications. At the same time only three correct classifications are taken out of the evaluation. In this way, the precision further increases (amounting to 100% in Figure 4d), while the recall still remains on a relatively high level (amounting to 60% in Figure 4d).

As the example described above shows, the cross-version voting approach succeeds in significantly increasing the precision, while keeping the recall at a relatively high level. For a quantitative evaluation of the cross-version voting strategy we refer to the experiments described in Section 3.3.

3.2 Constraint-Based Strategy

To better illustrate the potential of our cross-version voting strategy, we now consider a constraint-based stabilizing procedure. Using the template-based approach described in Section 2.3, the automatically derived chord label for a given bar is defined by the template having the minimal distance to the feature vector, in the following referred to as *basic strategy*. Figure 4e shows a visualization of the chord labeling result. As the visualization reveals the first bar is correctly identified as C♯ minor, whereas bar 2 is misclassified, being identified as E major although being labeled as C♯ minor in the ground truth. Here, a C♯ minor 7th chord is present in the ground truth, being mapped to C♯ minor. In fact, this seventh chord contains all the tones for E major, which explains the misclassification.

As we can see from the example, using the basic strategy, it obviously happens that for bars containing complex chords none of the given 24 templates fits well to the present feature vector. Here, the chord template of minimal distance may have a rather large distance to the feature vector. To counteract this case, we now introduce a parameter $\gamma \in [0, 1]$, which represents an upper threshold for the distance between the assigned chord template and the feature vector. In this way, we obtain a constraint-based procedure, where only chord labels λ are kept for which

$$d(\mathbf{t}_\lambda, x) < \gamma. \tag{2}$$

Table 1 Overview of the pieces and number of versions used in our experiments.

Composer	Piece	# (Versions)	Identifier
Bach	Prelude C Major BWV 846	5	'Bach'
Beethoven	Moonlight Sonata Op. 27 No. 2 (first movement)	7	'BeetM'
Beethoven	Fifth Symphony Op. 67 (first movement)	38	'Beet5'
Chopin	Mazurka Op. 68 No. 3	49	'Chopin'

All feature vectors x that have a larger distance than γ to any of the chord templates are left unannotated. In the following experiment, the idea is to successively decrease the parameter γ in order to investigate its influence on the chord labeling result.

Figure 4f shows the visualization for $\gamma = 0.3$. Obviously, one misclassification (bb. 16) is now taken out of the evaluation. However, at the same time two previously correctly classified chords (bb. 6, bb. 11) are left unconsidered in the evaluation, resulting in a decrease of the recall. Here, again seventh chords are present being correctly classified but having a relatively large distance to the template vector. Further decreasing the parameter γ is accompanied by a dramatical loss in recall while the precision increases moderately (Figure 4g). For quantitative results of the evaluation of the constraint-based strategy we refer to the experiments shown in Figure 5.

3.3 Experiments

In this section we quantitatively evaluate the various chord labeling strategies using a dataset that comprises four classical pieces of music, see Table 1. At this point, we want to emphasize that our main object is not in increasing the F-measure, defined below. Instead, in the application we have in mind, we are interested in finding passages, where one obtains correct chord labels with high guarantee. Therefore, our aim is to increase the precision, however, without losing too much of the recall.

In the following, we denote the automatically derived audio chord labels as L_a, and the ground truth chord labels as L_{gt}. For our bar-wise evaluation, we use precision (P), recall (R) and F-measure (F) defined as follows:

$$P = \frac{\#(L_a \cap L_{gt})}{\#L_a}, \quad R = \frac{\#(L_a \cap L_{gt})}{\#L_{gt}}, \quad F = 2 \cdot \frac{P \cdot R}{P + R}. \tag{3}$$

We first discuss the cross-version voting strategy. Figure 5 shows curves for P, R and F for the four pieces in the dataset, where the horizontal axis now represents the parameter ν ranging between 0.5 and 0.8 except for the position labeled by 'Maj' corresponding to the majority voting strategy. First of all, one notices that performing the chord labeling across several versions using the majority voting strategy, precision, recall and F-measure already improve by 10-30% in comparison to the basic strategy based on a specific version (see 'Min' in Figure 5).

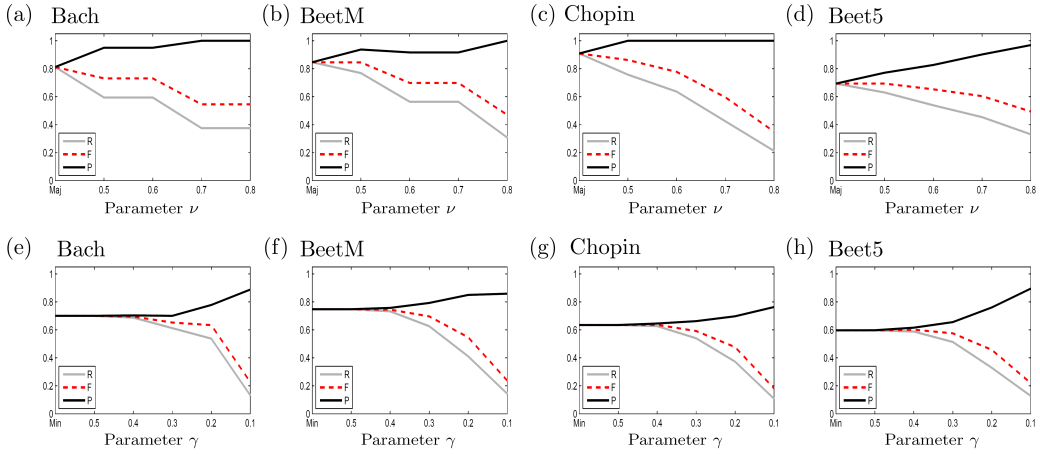


Figure 5 Top: Cross-version voting strategy. Curves for precision (P), recall (R) and F-measure (F) using the majority voting strategy (Maj), and four different consistency parameters ν from 0.5 to 0.8. **Bottom:** Constraint-based strategy based on a specific version. Curves for the mean value of precision (P), recall (R) and F-measure (F) using the basic strategy (Min) and five different settings for γ from 0.5 to 0.1.

Table 2 Basic chord labeling based on specific versions. The table shows mean, minimum and maximum F-measures over all recorded performances of a given piece.

	Mean	**Min**	**Max**
Bach	0.7000	0.4375	0.8750
BeetM	0.7473	0.6923	0.8718
Chopin	0.6345	0.4545	1.0000
Beet5	0.5967	0.5282	0.8345

Furthermore, for all four examples the precision rapidly increases, so that for $\nu = 0.5$ already a high precision is reached: 95% (Bach), 94% (BeetM), 100% (Chopin) and 77% (Beet5). At the same time the recall remains on a rather high level, still amounting to 59% (Bach), 77% (BeetM), 76% (Chopin) and 63% (Beet5). In this way, our experiments show that consistently labeled passages across several versions often correspond to correctly labeled passages. Increasing the consistency parameter ν further increases the precision values, while the recall still remains at acceptably high levels. In summary, exploiting the consistency information of the chord labels across several versions succeeds in stabilizing the chord labeling, resulting in a significant increase of precision without loosing too much of the recall.

We now compare these results with the ones obtained from the constraint-based strategy. Figure 5 shows curves for P, R, and F for the four pieces in our dataset. Here, P, R, and F correspond to mean values, which are obtained by first applying the constraint-based strategy on every version in the dataset separately and then averaging over all these versions. In the visualization, the horizontal axis represents the parameter γ ranging between 0.5 and 0.1 except for the position labeled by 'Min' corresponding to the basic labeling strategy. As one directly notices, there is a clear tendency visible for all four examples in our database. For increasing γ the precision also slowly increases reaching a high value of roughly 80% for

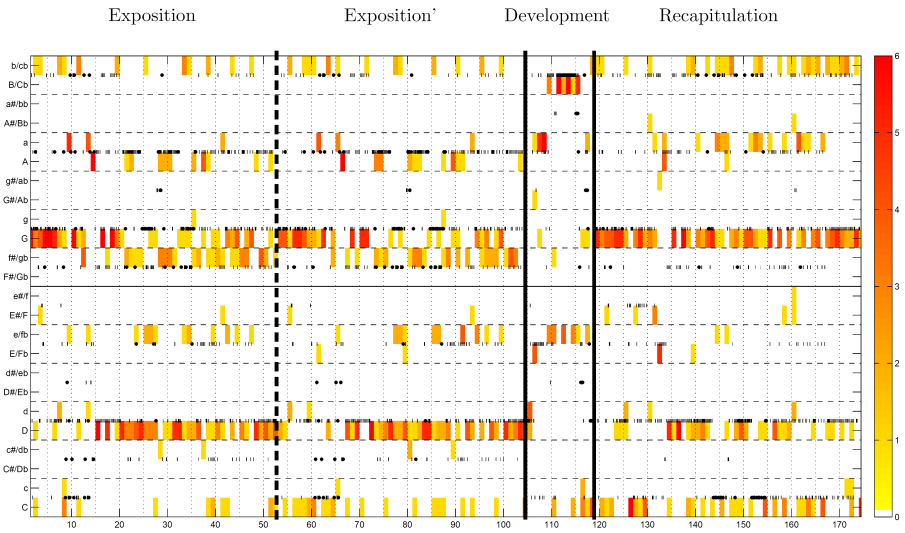


Figure 6 Cross-version visualization for the first movement of Beethoven's Piano Sonata Op. 49 No. 2. Here, six different recorded performances are considered.

$\gamma = 0.1$. However, at the same time the recall dramatically drops down to roughly 10% for $\gamma = 0.1$. Obviously, using the constraint-based strategy one can also increase precision values as misclassifications are taken out of the evaluation, however at the same time previously correct classifications are excluded resulting in a declining recall. Because of the dramatic loss of recall, this simple constraint-based strategy is not suited for stabilizing the chord labeling results.

Furthermore, our experiments reveal that performing the chord labeling based on a specific audio recording, the version-dependent results can vary greatly. This is shown by Table 2 indicating the mean F-measure, as well as the minimal and maximal F-measure achieved over all available recordings when using the basic labeling strategy (there was also a MIDI-synthesized version in each of the four groups). For example, the F-measure for one version of Bach amounts to 43.75%, corresponding to the minimal F-measure over all versions, whereas for another version the F-measure amounts to 87.5%, corresponding to the maximal F-measure over all versions. The average F-measure over the five versions amounts to 70%. These strong variations of the chord labeling results across different versions can not be explained by tuning effects, as we compensated for possible tuning deviations in the feature extraction step. A manual inspection showed that, for most cases, musical ambiguities are responsible for strong differences between the version-dependent results.

4 Exploring Harmonic Structures

As the experiments described above have shown, consistently labeled passages across several versions often correspond to correctly labeled passages. This opens the way for large-scale harmonic analyses on the basis of huge recorded music corpora, where cross-version chord labels of high reliability can be used instead of manually generated ground truth labels. In current work, we apply our cross-version analysis framework for automatically revealing hidden harmonic relations across different pieces of specific music corpora. In particular, in a collaboration with musicologists, we are investigating how to locate tonal centers, i. e.

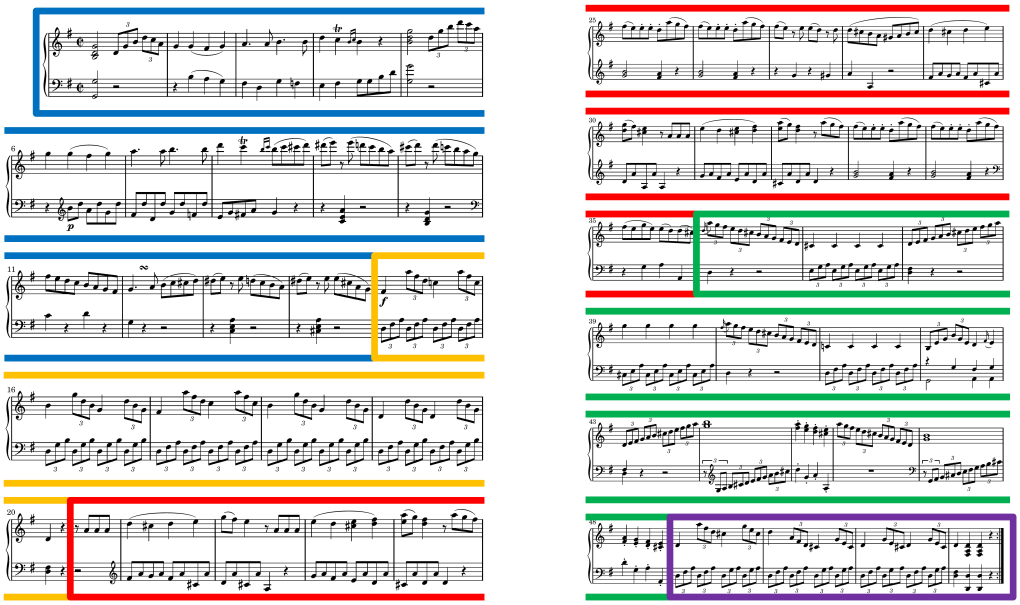


Figure 7 Exposition (bb. 1-52) of Beethoven's Piano Sonata Op. 49 No. 2. Musically meaningful sections are marked in the score: first group (blue), transition (yellow), second group (red), third theme (green), cadential group (purple).

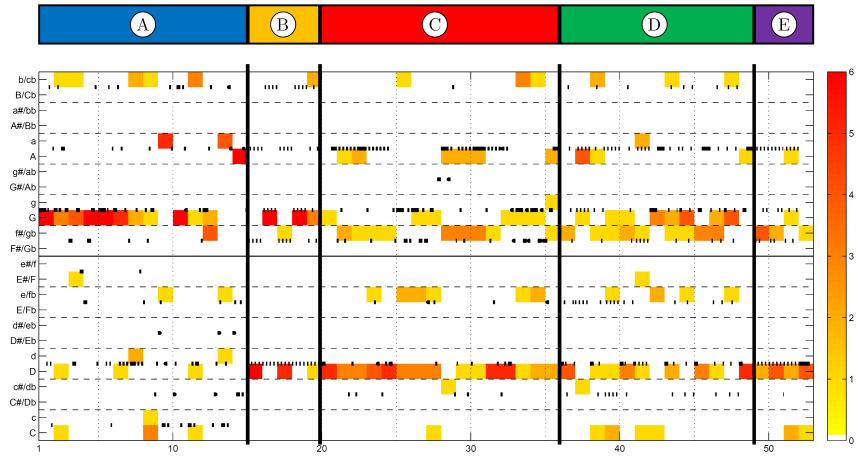


Figure 8 Cross-version visualization for the exposition of Beethoven's Piano Sonata Op. 49 No. 2. Here, six different recorded performances are considered.

passages which are dominated by a certain key, within large music corpora. Here, in the context of a harmonic analysis on a relatively coarse temporal level, our cross-version analysis has turned out to be a valuable tool that can reliably differentiate between harmonically stable and harmonically instable passages.

In the following, we exemplarily demonstrate how our cross-version visualization may serve musicologists as a helpful tool for exploring harmonic structures of a musical work. As example we use the first movement of Beethoven's Sonata Op. 49 No. 2 (see Figure 7). Figure 6 shows the cross-version visualization as an overview, where six different recorded performances are considered. The first movement is divided into three different form

parts: exposition (bb. 1-52) and its repetition (bb. 53-104), development (bb. 105-118) and recapitulation (bb. 119-174).[2] These parts are marked by vertical black lines and can be clearly separated from each other by their harmonic structures. The exposition is clearly dominated by the tonic G major and the dominant D major, which represent the keys of the first and the second theme, respectively. In contrast, the development is characterized by a greater variety of quickly changing harmonies: mainly D minor, A minor, E major, E minor and B major appear in the visualization as tonal centers. Finally, in the recapitulation, the tonic G major is stabilized: it appears now as the main tonal center (the second theme is likewise presented in the tonic), supported by shorter appearances of subdominant C major and dominant D major.

Apart from reflecting large-scale harmonic structures, the cross-version visualization allows for a more detailed bar-wise harmonic analysis, which we now exemplarily perform for the exposition of the sonata (see Figure 7 and Figure 8). The exposition is divided into five musically meaningful subparts, which again are characterized by specific harmonic structures: first group (A; bb. 1-14), transition (B; bb. 15-20), second group (C; bb. 20-35), third theme (D; bb. 36-48) and cadential group (E; bb. 49-52). These subparts of the exposition are marked by vertical black lines and displayed as color-coded blocks on top of the visualization (for a comparison to the score, see Figure 7).

As the visualization reveals the first theme (A) is characterized by harmonic stability, especially in the beginning where the tonic G major is clearly present. However, one directly observes two bars which are labeled inconsistently across the various performances indicating harmonic instability: bars 7 and 8. Comparing to the score, one finds out that in bar 7 indeed two different harmonies appear, which is the reason for the inconsistent labeling on the bar-level. Similarly, bar 8 contains several harmonies including a diminished seventh chord, a chromatic melodic line and a trill so that no unique chord label can be assigned to this bar. The transition (B) is characterized by harmonic stable passages in the tonic G major and the dominant D major. As the score reveals, this section is indeed characterized by a bar-wise change between these two chords so that the transition leads to the entrance of the second theme (C) appearing in the dominant D major. The visualization clearly reflects that the key of the second theme is D major. However, some of the bars also exhibit inconsistencies. For example, bars 21-24 are classified as F♯ minor instead of D major for some of the recordings. A closer look at the score reveals that in these introductory bars of the second theme the leading tone c♯ of D major is often present, which musically stabilizes D major but at the same time produces (together with the notes f♯ and a of the D major chord) a chord ambiguity leading to the classification F♯ minor for some of the performances. A similar confusion occurs in bars 28-30, where the performers strongly emphasize the leading tone c♯. The second theme is followed by a kind of third theme (D), which exhibits many inconsistencies. A comparison to the score shows that this passage is characterized by chromatic runs which are responsible for the inconsistent labeling across the considered performances. The exposition finally closes with the cadential group (E), which usually stabilizes the tonic in the end. Surprisingly, the visualization reveals that the labeling for this section is not as consistent as one may expect. Here, the score shows that the tonic D major indeed dominates this passage, but the leading tone c♯ appears again together with the suspended fourth g.

[2] Note, that bar numbering in printed sheet music usually does not take into account repetitions.

5 Conclusions

In this chapter, we presented a cross-version approach for chord labeling. In particular, we showed that consistently labeled passages across several versions often correspond to correctly labeled passages. Presenting the cross-version analysis results on a musically meaningful time axis in bars also helps to make the analysis results better accessible to music experts. Firstly, the presented approach allows for involving musicologists in the evaluation process of automated chord labeling procedures. For example, the cross-version visualization opens the way for an interdisciplinary collaboration, where musicologists may greatly support computer scientists in performing an in-depth error analysis of the employed chord labeling procedure based on the score. Secondly, the cross-version visualization may serve musicologists as a helpful tool for exploring harmonic structures of a musical work. Because of their high reliability, cross-version chord labels may be an alternative to manually generated ground truth labels. This may particularly hold for large-scale harmonic analyses on the basis of huge corpora of recorded music.

As for future work, we need to perform more detailed quantitative evaluations to verify our hypothesis that our cross-version approach indeed leads to a stabilization of the chord labeling results. Furthermore, we plan to apply our cross-version framework on the entire corpus of Beethoven's piano sonatas. In collaboration with musicologists, we are currently investigating harmonic structures across different movements for some of the sonatas. Here, our automated methods may help to investigate which tonal centers occur in a specific sonata and how they are functionally related to each other. In this context, a structure-oriented analysis, which analyzes tonal centers according to the different form parts of the classical sonata form, is of great musicological meaning as each such part is characterized by a specific occurrence of certain harmonies. Performing this analysis across the complete corpus of Beethoven's piano sonatas, we aim to quantify and better understand from a music-historical perspective how Beethoven has applied tonal centers in his work. Finally, we plan to use our automated framework for exploring harmonic structures across even larger and more complex corpora of musical works, such as the corpus of Wagner's operas. Here, due to the vast amount of data, a purely manual harmonic analysis is hardly possible. Also, being characterized by complex harmonies and rich orchestrations, the detection of large-scale harmonic relations within and across the operas becomes a challenging task.

6 Acknowledgment

We would like to express our gratitude to Michael Clausen, Cynthia Liem, and Matthias Mauch for their helpful and constructive feedback.

References

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

2 Juan Pablo Bello and Jeremy Pickens. A robust mid-level representation for harmonic content in music signals. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 304–311, London, UK, 2005.

3 Taemin Cho, Ron J. Weiss, and Juan Pablo Bello. Exploring common variations in state of the art chord recognition systems. In *Proceedings of the Sound and Music Computing Conference (SMC)*, pages 1–8, Barcelona, Spain, 2010.

4 Ching-Hua Chuan and Elaine Chew. Audio key finding: Considerations in system design and case studies on Chopin's 24 Preludes. *EURASIP Journal on Advances in Signal Processing*, 2007:1–15, 2007.

5 Sebastian Ewert, Meinard Müller, and Peter Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1869–1872, Taipei, Taiwan, 2009.

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

7 Emilia Gómez. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, 18(3):294–304, 2006.

8 Christopher Harte, Mark Sandler, Samer Abdallah, and Emilia Gómez. Symbolic representation of musical chords: A proposed syntax for text annotations. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 66–71, London, GB, 2005.

9 Ning Hu, Roger B. Dannenberg, and George Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, US, October 2003.

10 Verena Konz, Meinard Müller, and Sebastian Ewert. A multi-perspective evaluation framework for chord recognition. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, pages 9–14, Utrecht, The Netherlands, 2010.

11 Kyogu Lee and Malcolm Slaney. A unified system for chord transcription and key extraction using hidden Markov models. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Vienna, AT, 2007.

12 Robert Macrae and Simon Dixon. Guitar tab mining, analysis and ranking. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 453–458, Miami, USA, 2011.

13 Matthias Mauch and Simon Dixon. Simultaneous estimation of chords and musical context from audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1280–1289, 2010.

14 Matthias Mauch, Daniel Müllensiefen, Simon Dixon, and Geraint Wiggins. Can statistical language models be used for the analysis of harmonic progressions? In *Proceedings of the International Conference of Music Perception and Cognition (ICMPC)*, Sapporo, Japan, 2008.

15 Matt McVicar, Yizhao Ni, Raul Santos-Rodriguez, and Tijl De Bie. Using online chord databases to enhance chord recognition. *Journal of New Music Research*, 40(2):139–152, 2011.

16 MIREX 2010. Audio Chord Estimation Subtask. `http://www.music-ir.org/mirex/wiki/2010:Audio_Chord_Estimation`, Retrieved 17.09.2010.

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

18 Meinard Müller and Sebastian Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 215–220, Miami, USA, 2011.

19 Hélène Papadopoulos and Geoffroy Peeters. Joint estimation of chords and downbeats from an audio signal. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):138–152, 2011.

20 Jeremy T. Reed, Yushi Ueda, Sabato Siniscalchi, Yuki Uchiyama, Shigeki Sagayama, and Chin-Hui Lee. Minimum classification error training to improve isolated chord recognition. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 609–614, Kobe, Japan, 2009.

21 Craig Stuart Sapp. *Computational Methods for the Analysis of Musical Structure*. PhD thesis, Stanford University, USA, May 2011.

22 Alexander Sheh and Daniel P. W. Ellis. Chord segmentation and recognition using EM-trained hidden Markov models. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 185–191, Baltimore, USA, 2003.

23 Yushi Ueda, Yuuki Uchiyama, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama. HMM-based approach for automatic chord detection using refined acoustic features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5518–5521, Dallas, USA, 2010.

Score-Informed Source Separation for Music Signals

Sebastian Ewert*[1] and Meinard Müller†[2]

1 **Institute for Computer Science III, University of Bonn**
Römerstr. 164, 53117 Bonn, Germany
ewerts@iai.uni-bonn.de

2 **Saarland University and MPI Informatik**
Campus E1-4, 66123 Saarbrücken, Germany
meinard@mpi-inf.mpg.de

Abstract

In recent years, the processing of audio recordings by exploiting additional musical knowledge has turned out to be a promising research direction. In particular, additional note information as specified by a musical score or a MIDI file has been employed to support various audio processing tasks such as source separation, audio parameterization, performance analysis, or instrument equalization. In this contribution, we provide an overview of approaches for score-informed source separation and illustrate their potential by discussing innovative applications and interfaces. Additionally, to illustrate some basic principles behind these approaches, we demonstrate how score information can be integrated into the well-known non-negative matrix factorization (NMF) framework. Finally, we compare this approach to advanced methods based on parametric models.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems, I.5 Pattern Recognition

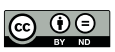
Keywords and phrases Audio processing, music signals, source separation, musical score, alignment, music synchronization, non-negative matrix factorization, parametric models

Digital Object Identifier 10.4230/DFU.Vol3.11041.73

1 Introduction

The decomposition of a mixture of superimposed acoustic sound sources into its constituent components, a task also known as *source separation*, is one of the central research topics in digital audio signal processing. For example, in speech signal processing, an important task is to separate the voice of a specific speaker from a mixture of conversations of multiple speakers and background noises ("Cocktail party scenario"), see for example [29]. Also in the field of musical signal processing, there are many related issues that are commonly subsumed under the notion of source separation. In the musical context, a source might correspond to a melody, a bassline, a drum track, or an instrument track. To extract such sources, various elaborate processing and analysis methods have been developed, which have led to significant improvements for tasks such as instrument recognition [22], harmonic analysis [47], or melody estimation [12]. Most of these methods exploit certain spectral and temporal

* Sebastian Ewert has been funded by the German Research Foundation (DFG CL 64/6-1).
† Meinard Müller has been supported by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI). He is now with Bonn University, Department of Computer Science III, Germany.



Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 73–94

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

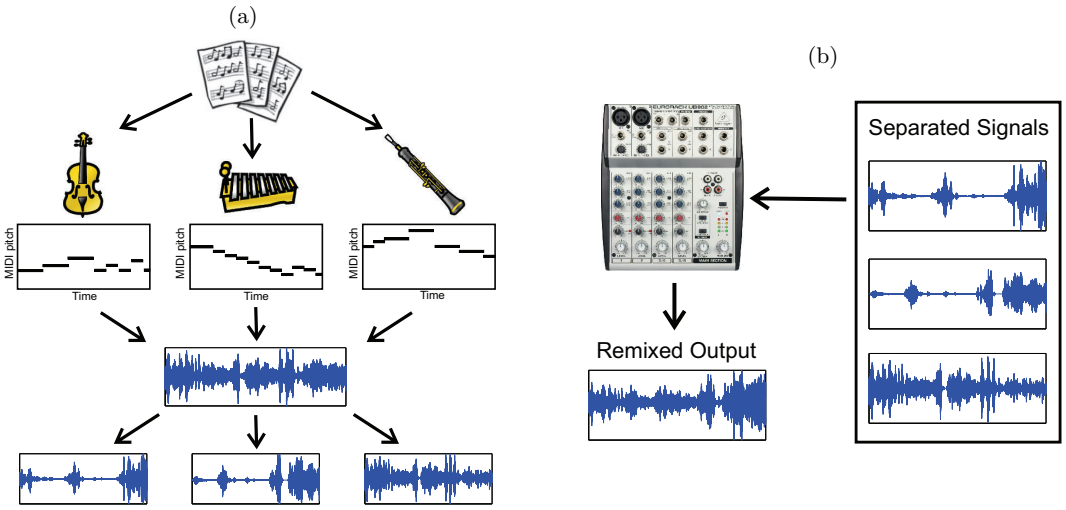


Figure 1 Score-informed source separation: **(a)** Instrument tracks as specified by a given score are employed for the separation of instrument sounds from a polyphonic audio recording (figure inspired by [24]). **(b)** Separated signals corresponding to instrument tracks can be remixed by the user in real-time (figure inspired by [27]).

properties of the sound sources to be extracted. For example, the melody is often the leading voice characterized by its dominance in dynamics and by its temporal continuity [3, 9]. The track of a bass guitar may be identified by specifically looking at the lower part of the frequency spectrum [19]. Furthermore, when extracting the drum track, one often relies on the assumption that the other sources are of harmonic nature. Then one can exploit that percussive elements (vertical spectral structures) are fundamentally different from harmonic elements (horizontal spectral structures) [36]. Last but not least, a human singing voice can often be distinguished from other musical sources because of the presence of vibrato and portamento (sliding voice) effects [40].

In the last years, also multimodal, score-informed source separation strategies have been employed where one assumes the availability of a score representation along with the music recording. The score provides valuable information in two respects. On the one hand, pitch and timing of note events provide a rough guidance within the separation process. On the other hand, the score provides a natural way of specifying what and how sound sources are to be separated. For example, in [24] the score's natural partition into instrument tracks is exploited to extract each individual instrument from a given audio recording, see Figure 1a for an illustration. Here, the score provides additional cues on the sources' spectral and temporal properties. In [27], it was demonstrated that this concept can be incorporated into an intuitive and easy-to-use interface. Here, the user can adjust the volume of each instrument in real-time using an interactive instrument equalizer, see Figure 1b. Developing this idea further, one can extend the instrument equalizer to a more general voice or note equalizer [14], where the user can not only emphasize or attenuate whole instrument tracks but also specific note groups played by different or the same instrument. Here, a group of notes might correspond to a motif, a voice, the left or the right hand of a piano score, or a staff as illustrated in Figure 2a. Incorporating these concepts into multimodal music players [5, 6], one can intuitively select note groups in the score and separate or enhance them in the audio recording in real-time, see Figure 2b.

In this contribution, we give an overview of strategies that employ score information

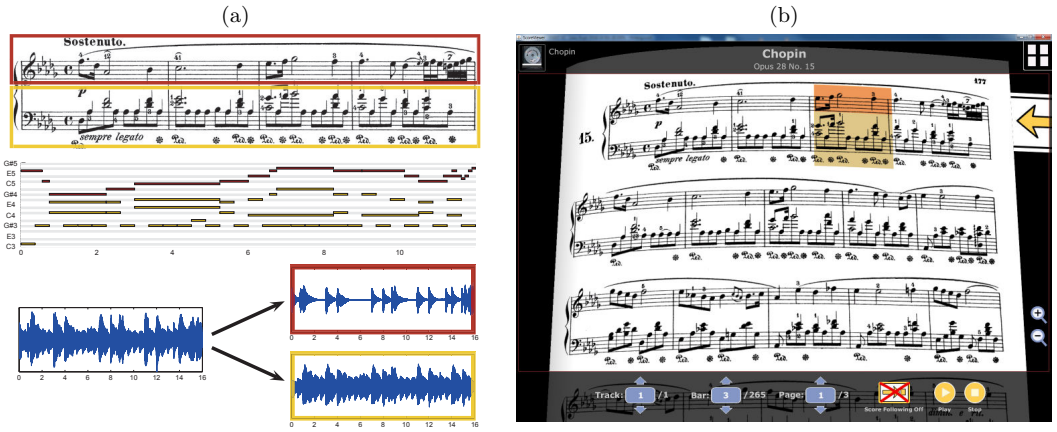


Figure 2 Score-informed voice separation: **(a)** Decomposition of a piano recording into two sound sources corresponding to the left and right hand as specified by a musical score. Shown are the first four measures of Chopin's Prélude "Raindrop" (Op. 28 No. 15). **(b)** Prototypical implementation of a voice equalizer based on the multimodal music player proposed in [5]. By selecting a staff/hand in the scanned score image the corresponding group of notes is separated/enhanced in real-time.

for separating musically meaningful sound sources from polyphonic music recordings. In Section 2, we summarize available score-informed source separation methods. Here, we focus on conceptual differences between the individual approaches rather than giving technical details. Then, using the well-known non-negative matrix factorization (NMF) framework as an example, we demonstrate in Section 3 how score information can be employed to guide the separation process. Finally, as an alternative to NMF-based approaches, we discuss in Section 4 advanced source separation methods based on parametric models. Conclusions and prospects on future work are given in Section 5.

2 Methods for Score-Informed Source Separation

In general, separating sound sources from polyphonic music recordings requires an understanding of many musical and technical aspects. For example, one has to account for the complexity of musical sound sources, the interaction and superposition of such sources in polyphonic mixtures, room acoustics, and recording conditions. Additionally, in many studio productions, numerous digital effect filters are applied to the recording thus making the task even more complex. However, although being extremely difficult, source separation is mostly pursued in a blind fashion, where as little prior knowledge as possible is used.

A natural idea to facilitate the separation process is to incorporate additional musical cues, for example, employing available musical score data. In this context, music synchronization methods are of particular importance [7, 8, 16, 28, 33]. Given a MIDI file representing the score and an audio recording representing an interpretation of a piece of music, the goal is to determine for each MIDI note event its corresponding time position in the audio recording. By adjusting the onset position and duration of each MIDI event, one can use the computed alignment to transform the original *score-like MIDI file* to a *synchronized MIDI file*, which runs synchronously to the audio, see Figure 3. Each score-informed source separation approach treats this problem differently. Some approaches consider or even account for typical differences between the score and a given interpretation, for example, in terms of structure, ornamentation, the interpretation of trills and arpeggios as well as additional

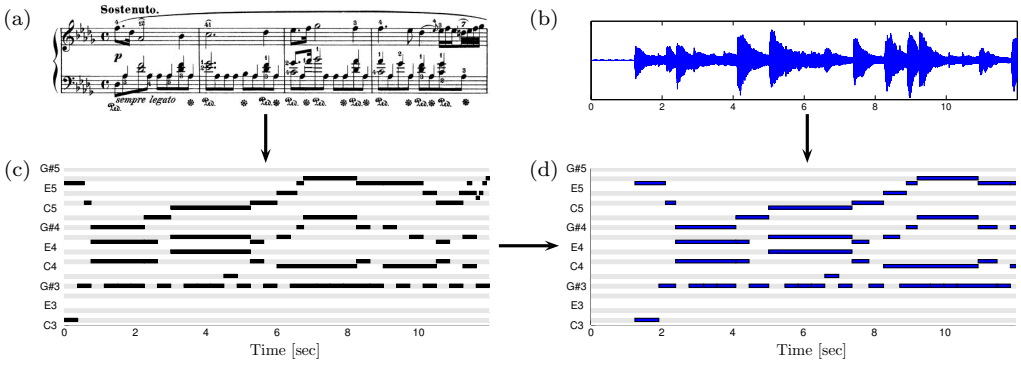


Figure 3 Music synchronization for a score and an audio recording of Chopin's Op. 28 No. 15: **(a)** Musical score. **(b)** Audio recording of an interpretation taken from the SMD database [34]. **(c)** Score-like MIDI file generated from the score shown in (a). **(d)** Synchronized MIDI file.

and missed notes. Other approaches simply assume that *perfectly synchronized MIDI files* are available. This assumption, however, is often not realistic. In real-world scenarios, one typically has to adjust a MIDI file to a given audio recording so that perfect synchronicity can not be guaranteed.

Early approaches adopt score and MIDI information only for evaluation purposes, for example, to investigate the influence of a pitch estimation step in a complex separation system [37]. One of the first approaches focusing on the conceptual benefits of incorporating score information was proposed in [42]. Here, the task consists in separating a single instrument specified by a given score-like MIDI file from a polyphonic music recording. The main idea is based on designing a filter, which in some sense optimally extracts the instrument from the recording. To compute the MIDI-audio synchronization, the authors refer to a procedure previously proposed in [41]. While presenting a novel application idea, this early work has several conceptual limitations. First of all, the proposed filter design procedure models all non-target sound sources as Gaussian noise. Therefore, in cases where the target instrument is accompanied by other instruments, this assumption is obviously violated. Furthermore, the proposed method assumes that the score provides an exact specification of the fundamental frequency for the target instrument for each analysis frame. This assumption is not realistic, since the score usually provides only high-level note information of the piece of music without specifying tuning or small pitch deviations of the respective music recording.

Subsequently proposed systems were not subject to such strict limitations. In [54], the authors integrate score information into a system for blind source separation previously described in [53] (an extended version was presented in [52]). Here, the goal is to extract individual instruments from a music recording, which then enables a user to create new music by remixing the extracted sound sources. In this approach, stereo information is employed in a first step to determine for each analysis frame the number of concurrent sources. Frames identified to contain only a single source are used as cues in the consecutive pitch-tracking step to support the separation in frames with multiple sources. The authors incorporate score information into this process as a rough guidance for the pitch-tracking. The underlying MIDI-audio alignment is based on a procedure proposed by Hu et al. [25]. A technical limitation of the approach is its dependency on reliable stereo information to identify the sources. This is problematic for many commercial studio productions, where spatial information contained in the stereo recordings is often corrupted by digital effect filters and virtual room acoustics. Furthermore, the influence of the alignment step is hard to

assess from the experimental results, as the method is only evaluated on a dataset consisting of four-second snippets of synthetically created MIDI sonifications.

While score information is used in [54] mostly as an add-on to an existing source separation system, Han and Raphael presented in [20, 21] a model that completely relies on available score data. In their contribution, the authors aim at removing the soloist from orchestral music recordings to generate recordings that can be used as a basis for automated accompaniment systems [7]. Relying on score information at an early stage of their algorithmic pipeline allowed for innovative computational concepts. On the one hand, the method represents a given input spectrogram as a compound of note-event based models. This allows for effectively using the score information to specify the temporal and spectral extent in which a note-event is permitted to be active. On the other hand, the score is used to identify the instruments occurring in a given music recording. This way, some instrument-dependent model parameters such as overtone energy distributions can simply be learnt from monophonic training material in advance and fixed afterwards. A benefit of this approach is that the parameter estimation process becomes efficient (as only a small set of parameters needs to be adjusted) and robust (as unreasonable parameter values are prevented by the model). However, a drawback is that the model can be imprecise, in particular when the training instruments differ strongly from the ones used in the given recording.

Roughly at the same time, Itoyama et al. presented a system, which explored novel application scenarios based on score-informed source separation [27]. This system allows a user to adjust the volume of each instrument in a polyphonic music recording in real-time. To this end, the system separates the individual instrument tracks in a preprocessing step as follows. In a first step, a MIDI synthesizer is employed to create one audio representation for each of the instrument tracks contained in a given MIDI file. This audio data is used as prior knowledge to initialize a note-based spectrogram model. Next, the model parameters are adapted to a given audio recording by minimizing a Kullback-Leibler distance between the given and the model spectrogram. Here, to allow only musically meaningful values for the model parameters, strong deviations form the initial values set in the first step are penalized. In a final step, the spectrogram model is employed to isolate the individual instrument tracks as specified by the MIDI file. Technically, the model is based on the harmonic-temporal-structured clustering (HTC) model proposed in [30], which will be discussed in more detail in Section 4. To control the influence of their percussion related submodel on the remaining system, the authors have to resort to smoothing and regulation techniques [26], which further increase the complexity of the system. Furthermore, alignment issues are not considered in this approach, hence it is not clear how the system behaves in real-world scenarios starting with score-like MIDI files.

Using MIDI-synthesized audio material for initialization purposes was also proposed by Gansemann et al. in [17, 18]. Given a MIDI file and an audio recording for a piece of music, the approach starts by sonifying the MIDI instrument tracks using a wavetable synthesizer similar to [27]. In a next step, probabilistic latent component analysis (PLCA) [43] is employed to identify the most important spectral components for each sonification. Here, PLCA is a probabilistic formulation of the well-known non-negative matrix factorization (NMF) method, which will be discussed in more detail in Section 3. In a last step, the instrument-wise spectral components are used as initialization and additional knowledge for a prior-based PLCA analysis of the original audio recording [45]. The results of this final analysis are subsequently used to extract each instrument from the original recording. Incorporating an alignment procedure by Turetsky and Ellis [46], the authors aim at using full-length score-like MIDI files as they can be found in real-world scenarios. While this approach presents a

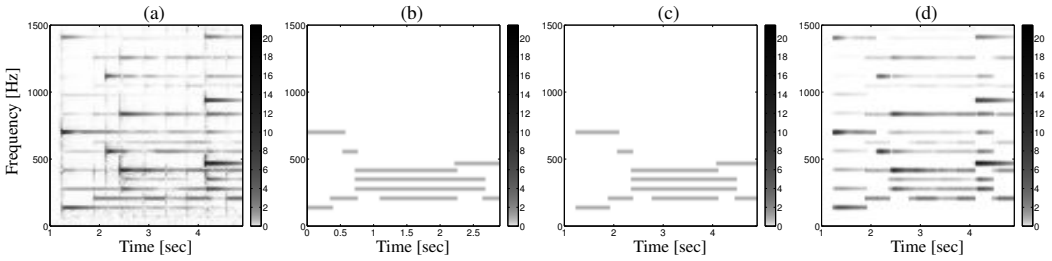


Figure 4 Score-informed parametric spectrogram model as employed in [13]. **(a)** Original magnitude spectrogram for a recording of Chopin's Op. 28 No. 15. **(b)** Model spectrogram after initialization with note events from a score-like MIDI file. **(c)** Model spectrogram after the synchronization step. **(d)** Model spectrogram after the estimation of remaining model parameters.

novel computational concept, the approach suffers from several weaknesses. Similar to all approaches relying on synthetic audio material as prior knowledge, this method's separation quality depends on the spectral similarity between the MIDI instruments and the actual target instruments. Moreover, this method also requires that the MIDI instruments have a similar tuning as the instruments in the given audio recording. For large tuning deviations, the separation quality might be significantly reduced.

An alternative way of using MIDI information for initialization purposes was presented in [24]. Here, instead of generating synthetic audio, the MIDI file is used to directly instruct the underlying spectrogram model when a given instrument is active with a certain pitch. This way, the separation performance does not depend on the quality of an underlying MIDI synthesizer. However, as a drawback, no expectations about the spectral shape of an instrument are incorporated, which may lead to a less robust separation process. As a novel contribution, the method employs a parametric NMF variant [23], which significantly enhances the modeling accuracy for instruments with vibrato and glissando. A technical limitation of this model is that all harmonic sounds in an analysis frame are assumed to be a compound of stationary sinusoidals. To evaluate the instrument separation quality of this approach, the authors neglect the alignment step and employ synthetic MIDI sonifications of Bach, Beethoven and Boccherini pieces.

While most score-informed source separation techniques aim at re-synthesizing the separation results with the goal to produce acoustically appealing sound sources, the method proposed in [13] employs these techniques for analysis purposes. Given a MIDI file and an audio recording for a piece of music, the task consists of estimating an intensity for each MIDI note event as occurring in the recording. On the one hand, this enables a user to analyze and compare different interpretations of a piece in terms of dynamics on a note-level. On the other hand, it allows for enriching a given score-like MIDI representation with performance-specific subtleties. The approach employs a parametric model that describes the spectrogram of a given recorded performance as a sum of note-event spectrograms, see Fig. 4. In a first step, the model is initialized with pitch, onset and duration information obtained from a given score-like MIDI file, see Fig. 4b. After that, music synchronization techniques are employed to determine for each note event the corresponding position in the audio recording, see Fig. 4c. In a next step, additional model parameters are iteratively refined such that the model spectrogram approximates the original spectrogram as accurately as possible, see Fig. 4d. In a final step, the individual note intensities are estimated using the adapted note-event spectrograms described by the model. The approach is evaluated based on audio and MIDI velocity values recorded via a Yamaha Disklavier. The influence of

the synchronization step is evaluated by artificially distorting the MIDI time information, which only roughly indicates the methods' performance for real-world score-like MIDI files.

As demonstrated in [14], a similar model can also be used to create acoustically appealing separation results. Here, the separation system is embedded into a multimodal music interface [6] to create a voice equalizer, see Figure 2b, which allows the user to intuitively select arbitrary note groups and attenuate or emphasize them in real-time. To demonstrate the applicability of this approach in real-world scenarios, the authors employ score-like MIDI files from the Mutopia Project[1] in combination with real audio recordings taken from the SMD [34] and European archive[2] databases and make their separation results available on a website[3]. One of the drawbacks of this system and the one proposed in [27] is that the separation has to be performed in advance, while the remixing step can be performed in real-time.

As demonstrated by Duan and Pardo in [10], the separation step can be performed in a low-delay real-time fashion. To this end, the authors replace the usually employed offline synchronization step by an online approach [11], which aligns a given MIDI file and a corresponding audio recording in real-time, a task often referred to as score-following [4, 7]. For each analysis frame, their separation system first estimates the exact fundamental frequency of each pitch using the aligned MIDI file as a guidance. In a next step, each pitch is extracted using a harmonic mask and assigned to one of the instruments as specified by the MIDI file. To make this process feasible in real-time, the mask is computed using a fixed overtone model, which is not adapted to a given recording.

Overall, while source separation has been a field of research for decades, using score information to guide the separation process is a relatively recent approach. As demonstrated by the contributions discussed in this section, score guidance allows for novel and innovative applications of source separation techniques. Furthermore, the additional musical cues provided by the score often allow for a gain in separation quality, which is difficult to achieve otherwise. Here, robust music synchronization techniques allow for using score-informed source separation methods in real-world scenarios, where usually no perfectly aligned MIDI file is available. In the next section, we give an impression of how score-informed source separation can be performed in practice.

3 Score-Informed Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF) has turned out to be a powerful tool for modeling, analyzing and separating the constituent parts of polyphonic music recordings. For example, NMF variants form the basis of methods for pitch estimation [2, 44], source separation [50], and pattern and motive identification [51]. However, using classic NMF it is often hard to predict which properties of the input are captured after the learning process. In this section, we show how the classical NMF framework can be extended in a straightforward way using available score data. As we will see, the basic idea is to replace the standard NMF initialization without changing the established and computationally efficient NMF learning process. This way, a musically meaningful factorization structure can be enforced, which stabilizes NMF-based source separation.

[1] http://www.mutopiaproject.org
[2] http://www.europarchive.org
[3] http://www.mpi-inf.mpg.de/resources/MIR/2011-ISMIR-VoiceSeparation/

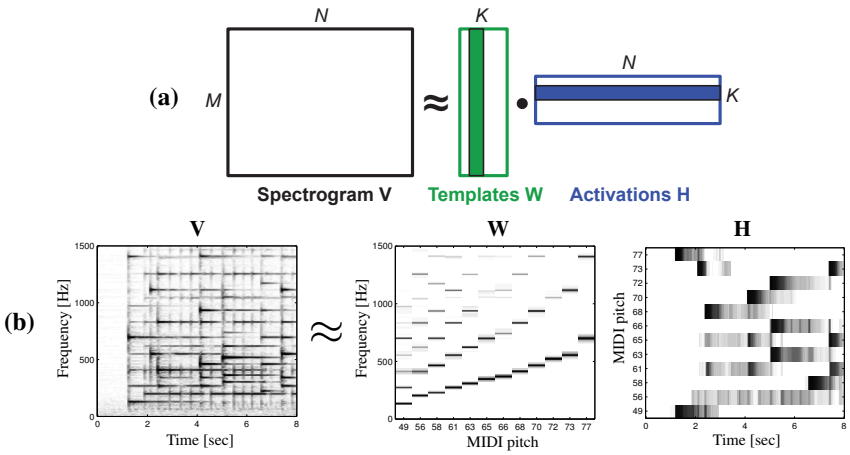


Figure 5 Non-negative matrix factorization (NMF). **(a)** A given non-negative matrix V is approximated as a product of two non-negative matrices W and H typically having a much smaller rank. **(b)** Example factorization of a magnitude spectrogram for an audio recording of Chopin's Op. 28 No. 15 taken from the SMD database [34].

3.1 Non-Negative Matrix Factorization

In classic non-negative matrix factorization, one approximates a spectral representation of a given recording by a product of two non-negative matrices. More exactly, given a magnitude spectrogram $V \in \mathbb{R}_{\geq 0}^{M \times N}$ of a music recording, NMF seeks to find non-negative matrices $W \in \mathbb{R}_{\geq 0}^{M \times K}$ and $H \in \mathbb{R}_{\geq 0}^{K \times N}$ such that $V \approx W \cdot H$, see Figure 5a. In this context, the columns of W are often referred to as *template vectors* and the rows of H as the corresponding *activations*. As an example, Figure 5b shows a factorization for a recording of Chopin's Op. 28 No. 15. Here, the free parameter K is set to the number of pitches that occur in the corresponding part of the piece. In this case, the activation matrix H is similar to a pianoroll representation and shows when these pitches become active.

In the classical approach for computing such a factorization, one employs some form of gradient descent to minimize a distance measure $D(V, W \cdot H)$ with respect to W and H, where D is typically based on the Euclidean norm or a variant of the Kullback-Leibler divergence, see [31]. However, to account for the non-negativity constraints for W and H, one usually has to resort to rather complex optimization algorithms [35]. As an easy-to-implement alternative, Lee and Seung proposed multiplicative update rules, which are derived from gradient descent by choosing a specific step size [31]. Using the popular Kullback-Leibler variant as a distance measure, these rules can be written as

$$H_{kn} \leftarrow H_{kn} \frac{\sum_i W_{ik} V_{in} / (WH)_{in}}{\sum_j W_{jk}} \quad \text{and} \quad W_{mk} \leftarrow W_{mk} \frac{\sum_i H_{ki} V_{mi} / (WH)_{mi}}{\sum_j H_{kj}},$$

where $m \in [1:M] := \{1, 2, \ldots, M\}$, $n \in [1:N]$, and $k \in [1:K]$. For vectorized programming languages such as Matlab it is useful to express these rules in matrix notation:

$$H \leftarrow H \odot \frac{W^\top \cdot (\frac{V}{W \cdot H})}{W^\top \cdot J} \quad \text{and} \quad W \leftarrow W \odot \frac{(\frac{V}{W \cdot H}) \cdot H^\top}{J \cdot H^\top},$$

where the $\cdot$ operator denotes the usual matrix product, the $\odot$ operator denotes the Hadamard product (point-wise multiplication), $J \in \mathbb{R}^{M \times N}$ denotes the matrix of ones, and the division is understood pointwise. These multiplicative update rules have several interesting

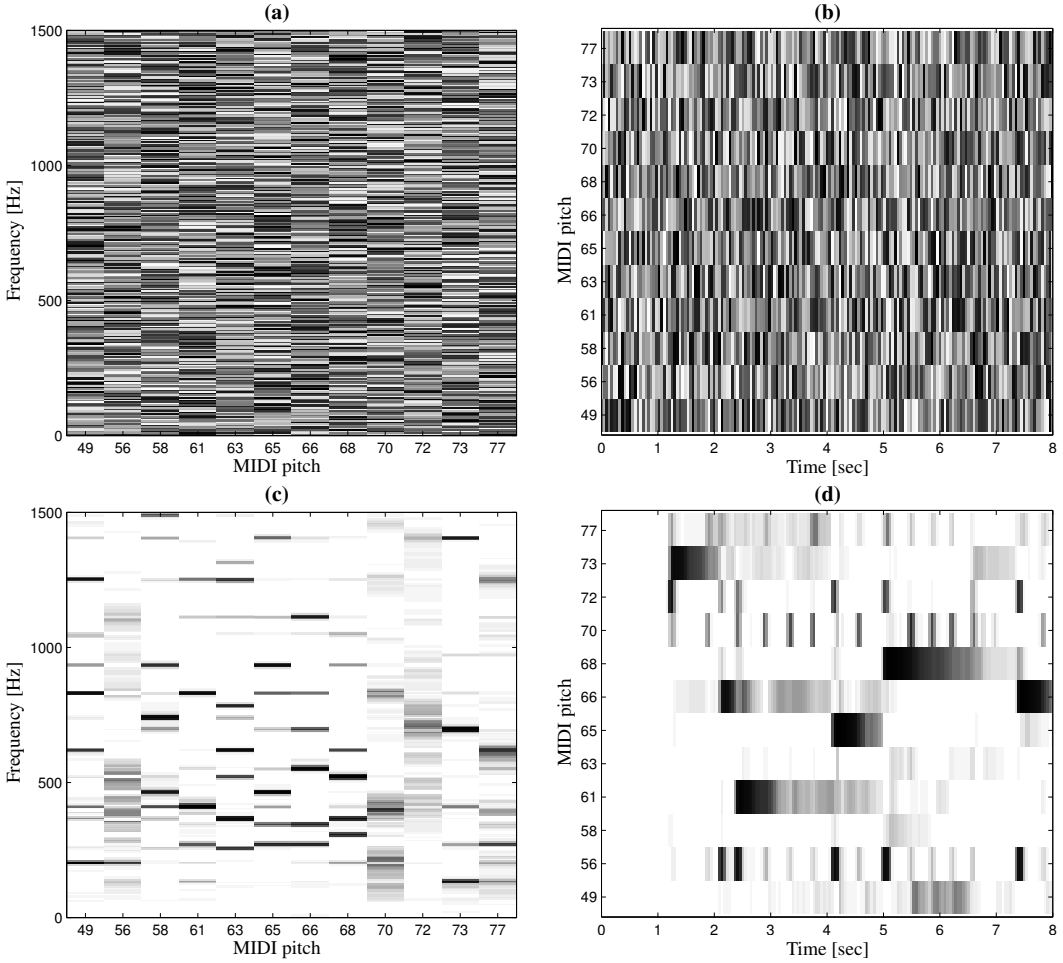


Figure 6 Classical NMF factorization for the magnitude spectrogram shown in Figure 5. **(a)** Random initialization of W. **(b)** Random initialization of H. **(c)** Learnt W. **(d)** Learnt H.

properties. First, the Kullback-Leibler distance measure is non-increasing under these rules[4]. Furthermore, initializing W and H with non-negative random values, these rules guarantee that W and H remain non-negative during the entire learning process.

In general, however, NMF factorizations computed in this classical way can not be as easily interpreted as the example shown in Figure 5b. For example, Figure 6 shows a factorization based on the classical NMF algorithm for the magnitude spectrogram shown in Figure 5b (again using $K = 12$). Here, the initialization of W and H with random values does not lead to a musically meaningful structure in the computed factorization. Furthermore, the free parameter K is usually set according to simple rules of thumb that usually do not account for any musical prior knowledge. As a result, the factorization often becomes completely unpredictable and lacks clear musical semantics.

Another important property of multiplicative update rules is that zero-valued entries remain zero during the entire learning process. Combined with musically informed initialization

[4] As pointed out by several authors [1, 32, 56], however, multiplicative rules do not guarantee in general convergence to a local minimum of the employed distance measure.

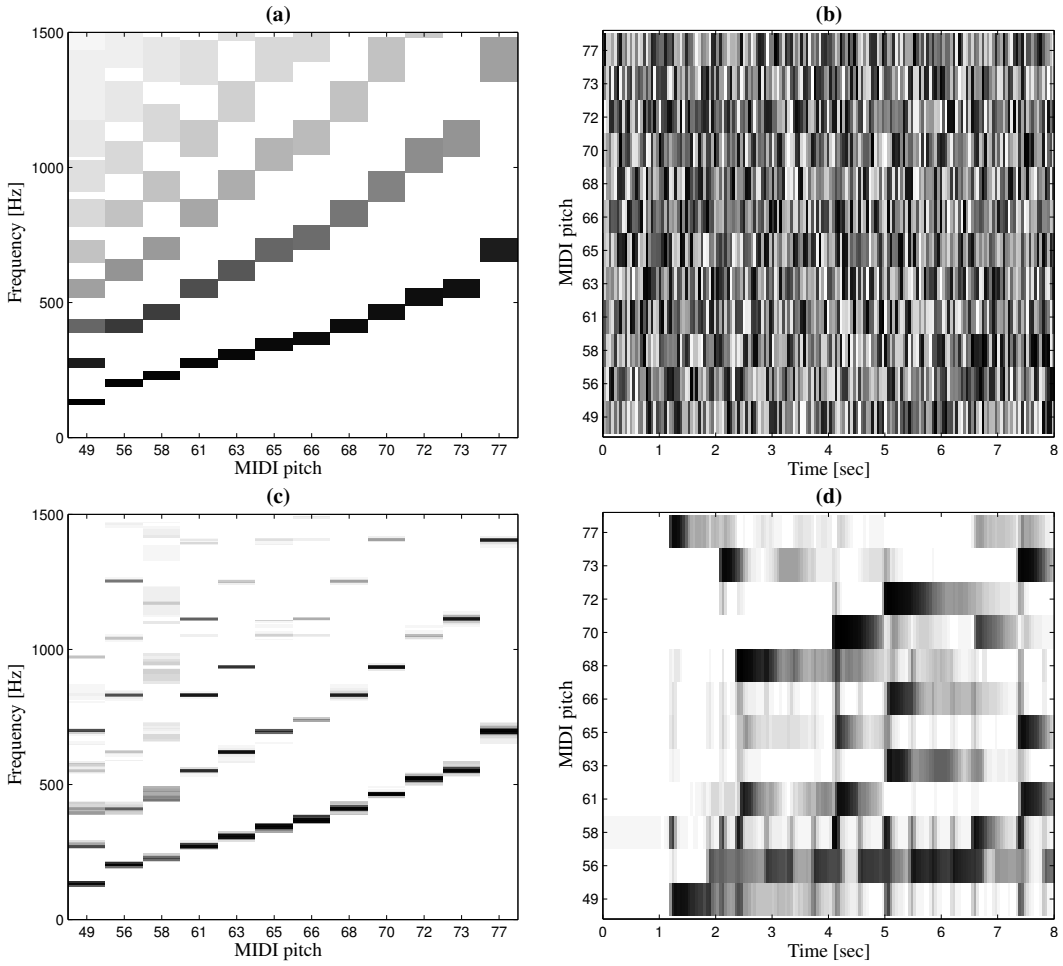


Figure 7 NMF factorization resulting from harmonic initialization of the template vectors for the magnitude spectrogram shown in Figure 5. **(a)** Harmonic initialization of W. **(b)** Random initialization of H. **(c)** Learnt W. **(d)** Learnt H.

schemes, this yields a straightforward way to enforce a specific structure of a factorization as proposed in [39, 49]. Here, one first creates one template vector for each possible MIDI pitch. Then, a harmonic structure is imposed by inserting zero-valued entries into the template initialization at positions where no partial is expected for a given pitch, see Figure 7a. The remaining entries are initialized according to a simplified overtone model. As we see in Figure 7c, the learning process based on multiplicative rules not only retains this harmonic structure but further refines it such that each template vector has a clear pitch association. This is a significant gain in structure compared to the unpredictable results computed via standard NMF as shown in Figure 6. However, looking at the resulting factorization in Figure 7c/d reveals that template vectors are still often 'misused', for example to represent onsets. This becomes particularly apparent in the template for MIDI pitch 58, where energy is distributed over a larger number of frequency bands compared to the other templates (Figure 7c). Here, instead of representing harmonic components of the spectrum, the template is misused to explain parts of the broadband energy distribution related to onsets. This is also reflected by the short-term intensity bursts in the corresponding activation row (Figure 7d).

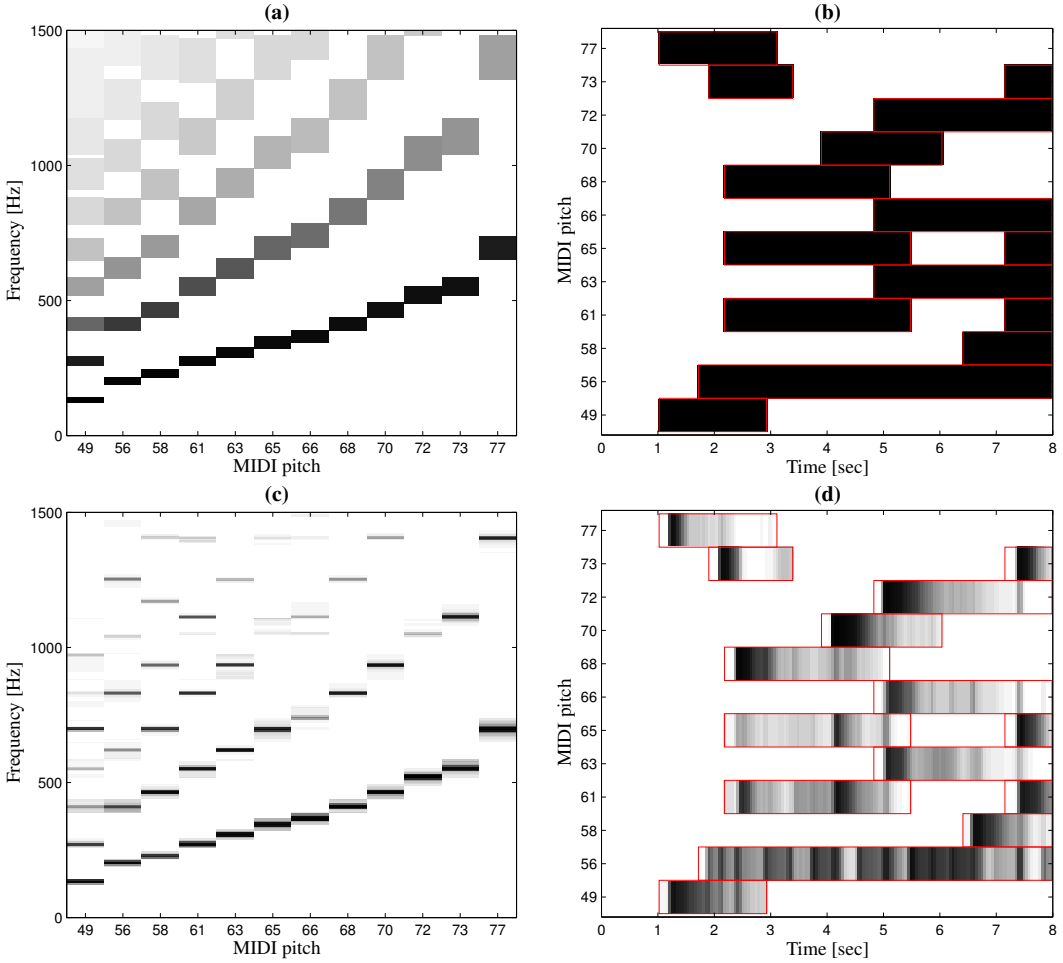


Figure 8 NMF factorization resulting from harmonic initialization of the template vectors and score-informed activation constraints for the magnitude spectrogram shown in Figure 5. **(a)** Harmonic initialization of W. **(b)** Score-informed initialization of H. **(c)** Learnt W. **(d)** Learnt H.

3.2 Integrating Score Information

Possible ways to further stabilize the factorization by incorporating additional score information were investigated in [15]. For example, in addition to the constraints on the template vectors, one can also impose constraints on the activations by incorporating note timing information. To generate such information, one employs music synchronization techniques in a first step to determine for each MIDI note event its corresponding position in the audio recording [16]. Next, based on the synchronized MIDI information, one marks suitable regions in H to determine where a given pitch can be active, see Figure 8b. The remaining entries are set to zero. To account for possible alignment inaccuracies, the temporal boundaries for these regions can be chosen relatively generous. As a result, the activation matrix H can be interpreted as a coarse piano roll representation of the synchronized MIDI file. As to be expected, combining these activation constraints with those for the template vectors further stabilizes the factorization. For example, most of the activation onset noise, which was present in Figure 7d, is suppressed in Figure 8d. Furthermore, almost all template vectors now have a well-defined harmonic structure. In some sense, the synchronization step can be

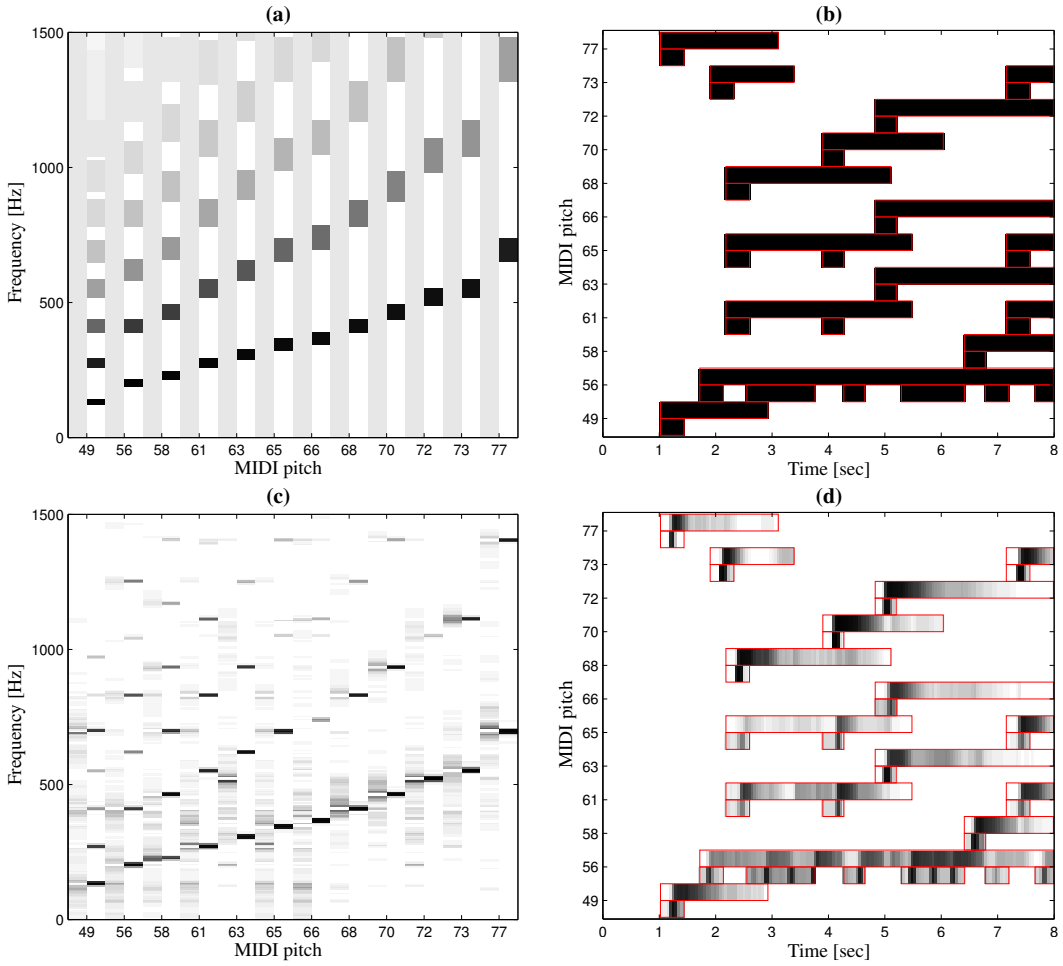


Figure 9 Extended NMF model with additional onset templates for the magnitude spectrogram shown in Figure 5. **(a)** Initialization of harmonic and onset template vectors in W. **(b)** Score-informed initialization of the corresponding activations in H. **(c)** Learnt W. **(d)** Learnt H.

seen to yield a first rough factorization, which is then refined by the NMF-based learning procedure.

So far, the model only represents harmonic parts of the signal and does not account for percussive elements such as onsets. Making again use of the score information, we extend the model by incorporating dedicated onset template vectors, see Figure 9a. Here, opposed to many other approaches, we take into account that the spectral shape for onsets is for many instruments (including the piano) not the same as for white noise but depends on the respective pitch. Therefore, instead of using one onset template jointly for all pitches as for example in [55], we use one onset vector for each pitch as suggested in [15]. Contrary to the harmonic templates, we do not enforce here any spectral constraints but initialize the onset templates uniformly and let the learning process derive their shape.

While the onset templates are hard to constrain in a meaningful way, the ephemeral nature of percussive sounds allows for imposing strict constraints on their activations. Using the synchronized MIDI, one has a rough estimate of the position of each onset. Initializing a small neighborhood around these positions to the value one in the corresponding activation while leaving all remaining entries at zero strongly restrains the time points where onset

templates are allowed to be active, see Figure 9b. Again, a tolerance should be used to compensate for possible synchronization inaccuracies. Looking at the resulting factorization shown in Figure 9c and 9d, we see that the learnt harmonic vectors have the clearest harmonic structure compared to all previous factorizations. Here, a reason is that percussive broadband energy is now captured by the onset templates, with the result that onsets now have a significantly less disturbing influence on the harmonic templates. Furthermore, the impulse-like activations of most onset templates at the start of note events indicate that these templates indeed represent onsets.

In summary, one can say that a combination of template and activation constraints leads to meaningful and robust matrix factorizations. Here, as for the case of the onset templates, constraints on the activation side can compensate for using relatively loose or even no constraints on the template side and vice versa. Furthermore, even though all constraints are hard in the sense that zero-entries in W and H remain zero throughout the learning process, one can use rather generous constraint regions to account for synchronization errors and retain some degree of flexibility. As one major advantage, the extended NMF model using hard constraints allows for using exactly the same multiplicative update rules as in classical NMF, thus it inherits the ease of implementability and computational efficiency.

3.3 Separation Process

By means of the initial constraint regions, a factorization as shown in Figure 9 describes how each note event of a given MIDI file manifests in the spectrogram of a corresponding audio recording. We now describe how this spectrogram model can be employed to separate note groups such as a melody line, the staff of the right hand, a specific motive, or the accompaniment from the recording. The only requirement is that the notes to be considered are somehow specified by the user or by some labeling of the score. As an illustrating example, we consider here the task of separating the left from the right hand staff as specified by a given score, see Figure 10a. While staffs do not always correspond to musically meaningful note groups, it demonstrates how note groups could be easily specified in a natural way.

For the separation, we exploit that every non-zero entry in H is associated with a specific note event, see Figure 9d. Therefore, we can partition H into two new matrices H_L and H_R, which contain either the activations for the left or the right hand, see Figure 10a/b. A straightforward way to create an audible separation result could be to multiply these two matrices with the template matrix W, shown in Figure 9c, and to invert the resulting spectrogram. However, as NMF-based models are typically used to compute a rough approximation of the original magnitude spectrogram spectral nuances in a given recording are usually not captured. Therefore, the resulting audio recording would sound rather unnatural.

An alternative to this direct sonification is commonly referred to as masking. Here, one first derives masking matrices via

$$M_L := \frac{WH_L}{WH+\varepsilon} \qquad \text{and} \qquad M_R := \frac{WH_R}{WH+\varepsilon},$$

where the division is understood pointwise and ε is a small positive constant to avoid a potential division by zero, see Figure 10c/d. M_L and M_R have the same size as the original spectrogram V and, having values between 0 and 1, indicate how strongly each entry in V belongs to either the left or the right hand. Multiplying these masking matrices point-wisely with V, one obtains a separated spectrogram for the left and the right hand, see Figure 10e/f. Finally, to obtain the separated audio signals, one applies an inverse

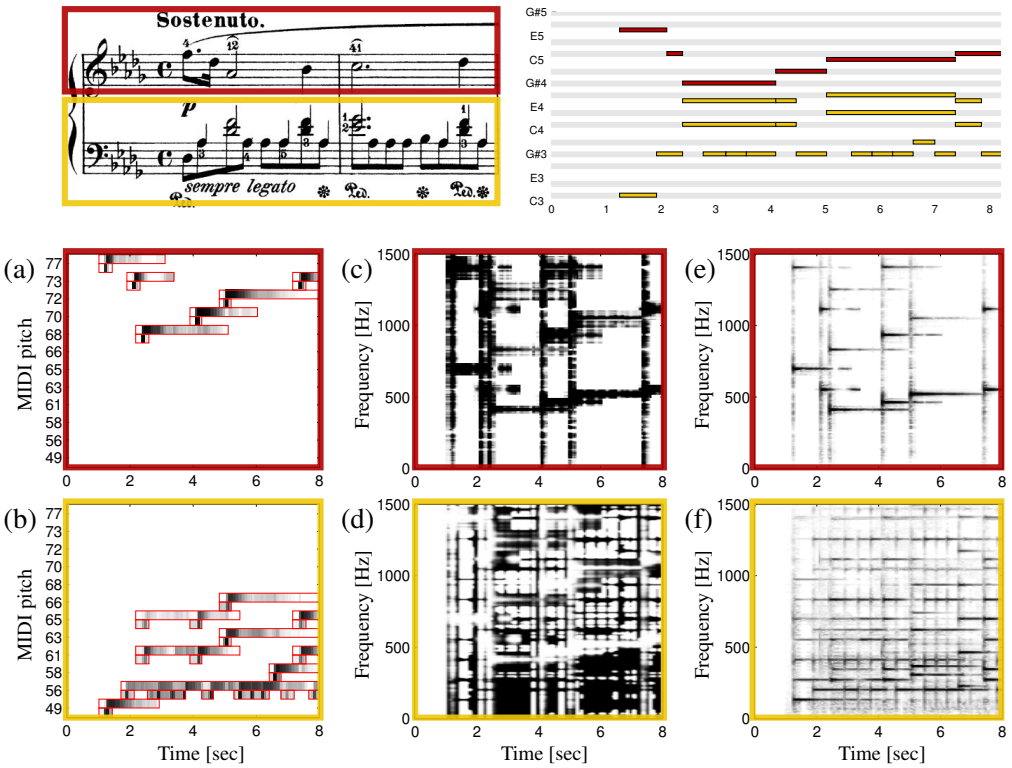


Figure 10 Illustration of the separation process for the left and the right hand. **(a)**/**(b)**: Partition of the activation matrix H (Figure 9d) into H_L and H_R. **(c)**/**(d)**: Masking matrices M_L and M_R. **(e)**/**(f)**: Separated spectrograms.

discrete Fourier transform in combination with an overlapp-add technique to the separated spectrograms. The necessary phase information is provided by the original spectrogram. This way, masking-based separation allows for preserving most spectral details of the original recording, which is important to create acoustically appealing results. However, by filtering the original audio data, masking may also retain more non-target spectrogram components compared to a direct sonification.

The quality of a separation result is often measured in terms of signal-to-distortion ratios (SDR) as proposed in [48]. While illustrating some general tendencies, these measures often do not capture the overall perceptual separation quality. In particular, in combination with synthetic audio material, one does not get an impression of the separation quality in real-world scenarios. To allow for a subjective, perceptual evaluation of their score-informed NMF variant, the authors in [15] provide a website[5] with separation results using real audio recordings and score-like MIDI files. Here, using full-length pieces by Bach, Beethoven and Chopin, most the audio material was taken from the SMD [34] dataset, while the MIDI files were provided by the Mutopia Project[6]. Some additional historical recordings were also taken from the European Archive[7]. To roughly indicate general quality differences between the NMF variants in a quantitative fashion, the authors also conducted experiments

[5] `http://www.mpi-inf.mpg.de/resources/MIR/ICASSP2012-ScoreInformedNMF/`
[6] `http://www.mutopiaproject.org`
[7] `http://www.europarchive.org`

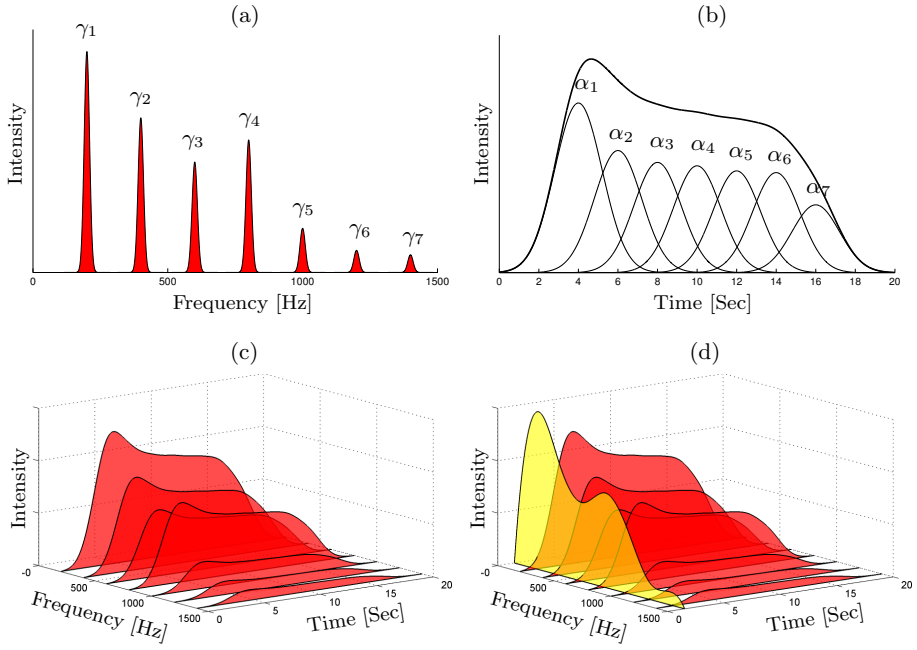


Figure 11 Harmonic-Temporal-Structured Clustering (HTC): **(a)** Template vector composed of several Gaussians. **(b)** Activation described by smooth, overlapping Gaussians. **(c)** Spectrogram model resulting from a combination of template vectors and activations similar to NMF. **(d)** Advanced HTC variant with an additional transient submodel. Figures are inspired by [55].

based on synthetic audio and the SDR measure. Here, on average, the strategy based on the harmonic initialization of W yielded the lowest SDR value. Combining this strategy with the score-informed initialization of H as in Figure 8 leads to a significant SDR-gain of roughly 1.5 dB. Finally, additionally integrating onset templates leads to another gain of roughly 1.2 dB.

4 Parametric Models

In addition to NMF, there are numerous other classical source separation methods which allow for the integration of score information. Many of the approaches discussed in Section 2 are based on so called *parametric models* [13, 14, 20, 24, 27], which have been widely used for blind source separation and music transcription. While these approaches differ strongly in their details, the common idea is to adapt a set of parameters such that the underlying model explains the spectrogram of a given recording as accurately as possible. Here, typical parameters are related to acoustical and musical properties such as pitch, amplitude and timbre. In this section, we exemplarily discuss some aspects of the *harmonic-temporal-structured clustering model (HTC)* [55], which was employed in [26, 27] for score-informed source separation. After a brief description of the main ideas underlying the HTC approach, we summarize some conceptual differences to the NMF model.

4.1 Harmonic-Temporal-Structured Clustering (HTC)

Harmonic-Temporal-Structured Clustering (HTC) employs a parametric model to approximate the magnitude spectrogram of a given audio recording. Compared to NMF, specialized

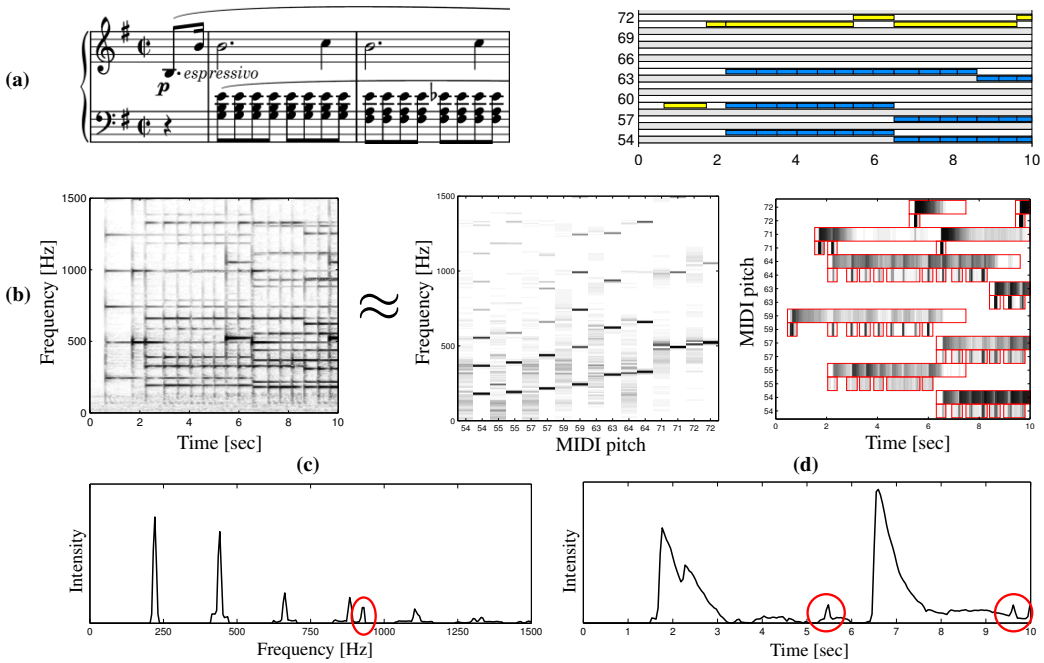


Figure 12 Score-informed NMF factorization for a recording of Chopin's Op. 28 No. 4 taken from the SMD database [34]. **(a)** Score and MIDI representation. **(b)** NMF factorization computed using the method presented in Section 3. **(c)** Zoomed template vector for pitch 57. **(d)** Zoomed activation for pitch 71. The red markers indicate positions discussed in the text.

model components take over the role of template vectors and activations. For example, each HTC template consists of several Gaussians, which represent the partials of a harmonic sound, see Figure 11a. To adapt the model to different instruments and their specific overtone energy distribution, the HTC model allows for scaling the height of each Gaussian individually using a set of parameters ($\gamma_1, \ldots, \gamma_7$ in Figure 11a). An additional parameter f_0 specifies the fundamental frequency for the template. Assuming a harmonic relationship between the overtones, this parameter controls the exact location of each partial.

Gaussians are also used in HTC to represent the activations, see Figure 11b. The position of these Gaussians is typically fixed such that only suitable height parameters can be adapted (parameters $\alpha_1, \ldots, \alpha_7$ in Figure 11b). By choosing suitable values for the variances and positions of the Gaussians, one obtains an overall smooth activation progression. Combining the HTC templates and activations similar to NMF, one obtains a spectrogram model as shown in Figure 11c. Recently proposed extensions of this model even allow for an integration of transient and onset models [27, 55], see Figure 11d. Again using a smoothed representation based on Gaussians, these additional models represent the broadband energy distribution usually found at onset positions.

Since the HTC model follows similar ideas as NMF, one can also employ similar strategies to incorporate score information. For example, note timing information can be used to restrict the use of the activation parameters. Furthermore, MIDI pitch information can be used to set the number of templates in the HTC model to the actual number of pitches in the piece. This is similar to setting the value of the free parameter K in NMF.

4.2 Comparison between HTC and NMF

To compare the HTC model with NMF, we consider an NMF factorization for an audio recording of Chopin's Prélude No. 4. Using the method presented in Section 3, one obtains a factorization as shown in Figure 12b. Here, similar to Figure 9, we see that almost all learnt harmonic template vectors have a well-defined harmonic structure. For a closer inspection of an exception, we plot the template for pitch 57 as a function over frequency in Figure 12c. We see a small peak at 930 Hz (see red marking), which does not fit into the harmonic pattern. Enforcing a meaningful distance between partials, the Gaussian-based HTC model offers here a straightforward way to enforce a clear harmonic relationship. However, this additional robustness against spurious peaks comes at the cost of model inaccuracies. One reason is that partials almost never perfectly take the form of a Gaussian, see [38], such that the HTC model leads to an additional inevitable approximation error.

Furthermore, the approximation accuracy does not only depend on the templates but also on the activations. To give an example, we plot the activation for pitch 71 as a function over time in Figure 12d. Here, we see three distinct peaks at 1.8, 2.3 and 6.6 seconds, respectively, which correspond to the three middle B notes, see Figure 12a. However, there are additional, smaller peaks at 5.4 and 9.6 seconds (marked in red), which do not seem to make any musical sense. Using Gaussians spanning several frames to model the activation, such short-time irregular peaks are smoothed out. However, whether this is meaningful depends on the application. In Figure 12a, we see that a note event with pitch 71 (middle B) is played after 2.3 seconds and is held afterwards. Then, after 5.4 seconds, a note event with pitch 72 is played. Since in this recording all piano dampers are up, the consequence is that the onset of pitch 72 also results in excitations of the neighboring pitches, in particular of the strings of pitch 71. Therefore, the small peak at 5.4 seconds in the NMF activation is indeed a physical fact rather than an extraction error.

5 Conclusion

Music signals possess specific characteristics that are not shared by spoken speech or audio signals from other domains. For example, for sound mixtures of polyphonic music, the general assumption that sources are somehow orthogonal in the spectral domain is often violated. This makes the separation of musical sources or voices very difficult. To remedy this problem, various approaches have been suggested that use additional cues as specified by a musical score.

In this paper, we have given a comprehensive overview of state-of-the-art source separation techniques that exploit additional score information in various ways. In particular, we discussed in detail a score-informed variant of NMF, where the integration of constraints can be done in a straightforward manner already at the initialization stage. We showed that by constraining both the template vectors and the activations, one obtains robust and musically meaningful separation results. Opposed to parametric models, where the integration of additional priors often leads to an increase in the computational complexity, score-informed NMF variants employ the same update rules as the original NMF and inherit its computational efficiency.

Besides stabilizing the separation process, the availability of score information also facilitates a natural and user-friendly way of specifying the voices or note groups to be separated. This opens up new ways for audio editing applications, where a user can simply mark certain note groups within a visual representation of the score, which are then separated, removed, amplified, or attenuated in a corresponding music recording. For the future, we plan to develop multimodal interfaces that realize such functionalities.

So far, we have conducted experiments mainly on piano music. In this context, we showed how the score-informed NMF framework can be extended by integrating additional onset templates without sacrificing robustness. A promising research direction is to further expand the NMF model to account for other musical aspects such as timbre or instrumentation and then to apply the NMF framework to other types of music.

6 Acknowledgment

We would like to express our gratitude to Björn Schuller, Joan Serrà, and Steve Tjoa for their helpful and constructive feedback.

References

1 Roland Badeau, Nancy Bertin, and Emmanuel Vincent. Stability analysis of multiplicative update algorithms for non-negative matrix factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2148–2151, Prague, Czech Republic, 2011.

2 Nancy Bertin, Roland Badeau, and Emmanuel Vincent. Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):538–549, 2010.

3 Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

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

5 David Damm, Christian Fremerey, Frank Kurth, Meinard Müller, and Michael Clausen. Multimodal presentation and browsing of music. In *Proceedings of the 10th International Conference on Multimodal Interfaces (ICMI)*, pages 205–208, Chania, Crete, Greece, 2008.

6 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: Special Issue on Music Digital Libraries*, 2011, to appear.

7 Roger B. Dannenberg and Christopher Raphael. Music score alignment and computer accompaniment. *Communications of the ACM, Special Issue: Music information retrieval*, 49(8):38–43, 2006.

8 Simon Dixon and Gerhard Widmer. MATCH: A music alignment tool chest. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, GB, 2005.

9 Karin Dressler. An auditory streaming approach for melody extraction from polyphonic music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 19–24, Miami, USA, 2011.

10 Zhiyao Duan and Bryan Pardo. Soundprism: An online system for score-informed source separation of music audio. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1205–1215, 2011.

11 Zhiyao Duan and Bryan Pardo. A state space model for online polyphonic audio-score alignment. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 197–200, Prague, Czech Republic, 2011.

12 Jean-Louis Durrieu, Gaël Richard, Bertrand David, and Cédric Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):564–575, 2010.

13 Sebastian Ewert and Meinard Müller. Estimating note intensities in music recordings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 385–388, Prague, Czech Republic, 2011.

14 Sebastian Ewert and Meinard Müller. Score-informed voice separation for piano recordings. In *Proceedings of the International Society for Music Information Retrieval Conference(ISMIR)*, pages 245–250, Miami, USA, 2011.

15 Sebastian Ewert and Meinard Müller. Using score-informed constraints for NMF-based source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

16 Sebastian Ewert, Meinard Müller, and Peter Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1869–1872, Taipei, Taiwan, 2009.

17 Joachim Ganseman, Paul Scheunders, Gautham J. Mysore, and Jonathan S. Abel. Evaluation of a score-informed source separation system. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 219–224, Utrecht, The Netherlands, 2010.

18 Joachim Ganseman, Paul Scheunders, Gautham J. Mysore, and Jonathan S. Abel. Source separation by score synthesis. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 462–465, New York, USA, 2010.

19 Masataka Goto. A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication (ISCA Journal)*, 43(4):311–329, 2004.

20 Yushen Han and Christopher Raphael. Desoloing monaural audio using mixture models. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 145–148, Vienna, Austria, 2007.

21 Yushen Han and Christopher Raphael. Informed source separation of orchestra and soloist. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 315–320, Utrecht, The Netherlands, 2010.

22 Toni Heittola, Anssi P. Klapuri, and Tuomas Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 327–332, Kobe, Japan, 2009.

23 Romain Hennequin, Roland Badeau, and Bertrand David. Time-dependent parametric and harmonic templates in non-negative matrix factorization. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 246–253, Graz, Austria, 2010.

24 Romain Hennequin, Bertrand David, and Roland Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 45–48, Prague, Czech Republic, 2011.

25 Ning Hu, Roger B. Dannenberg, and George Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2003.

26 Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Integration and adaptation of harmonic and inharmonic models for separating polyphonic musical signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages I–57–I–60, Hawaii, USA, 2007.

27 Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Instrument equalizer for query-by-example retrieval: Improving sound source separ-

ation based on integrated harmonic and inharmonic models. In *Proceedings of the International Conference for Music Information Retrieval (ISMIR)*, pages 133–138, Philadelphia, USA, 2008.

28 Cyril Joder, Slim Essid, and Gaël Richard. A comparative study of tonal acoustic features for a symbolic level music-to-score alignment. In *Proceedings of the 35nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, USA, 2010.

29 Cyril Joder, Felix Weninger, Florian Eyben, David Virette, and Björn Schuller. Real-time speech separation by semi-supervised nonnegative matrix factorization. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel Aviv, Israel, 2012.

30 Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama. Harmonic-temporal-structured clustering via deterministic annealing EM algorithm for audio feature extraction. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 115–122, London, GB, 2005.

31 Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 556–562, Denver, CO, USA, 2000.

32 Chih-Jen Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 18:1589–1596, 2007.

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

34 Meinard Müller, Verena Konz, Wolfgang Bogler, and Vlora Arifi-Müller. Saarland music data (SMD). In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR): Late Breaking session*, 2011.

35 Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer (Springer Series in Operations Research and Financial Engineering), 2006.

36 Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka, and Shigeki Sagayama. A real-time equalizer of harmonic and percussive components in music signals. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 139–144, Philadelphia, Pennsylvania, USA, 2008.

37 Mark D. Plumbley, Samer A. Abdallah, Juan Pablo Bello, Mike E. Davies, Giuliano Monti, and Mark B. Sandler. Automatic music transcription and audio source separation. *Cybernetics and Systems*, 33(6):603–627, 2002.

38 John G. Proakis and Dimitris G. Manolakis. *Digital Signal Processsing*. Prentice Hall, 1996.

39 Stanislaw Andrzej Raczynski, Nobutaka Ono, and Shigeki Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 381–386, 2007.

40 Lise Regnier and Geoffroy Peeters. Singing voice detection in music tracks using direct voice vibrato detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1685–1688, Taipei, Taiwan, 2009.

41 Shai Shalev-Shwartz, Shlomo Dubnov, Nir Friedman, and Yoram Singer. Robust temporal and spectral modeling for query by melody. In *Proceeding of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 331–338, Tampere, Finland, 2002.

42 Adiel Ben Shalom, Shai Shalev-Shwartz, Michael Werman, and Shlomo Dubnov. Optimal filtering of an instrument sound in a mixed recording using harmonic model and score alignment. In *Proceedings of the International Computer Music Conference (ICMC)*, Miami, USA, 2004.

43 Madhusudana Shashanka, Bhiksha Raj, and Paris Smaragdis. Probabilistic latent variable models as nonnegative factorizations (article id 947438). *Computational Intelligence and Neuroscience*, 2008.

44 Paris Smaragdis and Judith C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 177–180, 2003.

45 Paris Smaragdis and Gautham J. Mysore. Separation by humming: User guided sound extraction from monophonic mixtures. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 69–72, New Paltz, NY, USA, 2009.

46 Robert J. Turetsky and Daniel P.W. Ellis. Ground-truth transcriptions of real music from force-aligned MIDI syntheses. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 135–141, Baltimore, USA, 2003.

47 Yushi Ueda, Yuuki Uchiyama, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama. HMM-based approach for automatic chord detection using refined acoustic features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5518–5521, Dallas, USA, 2010.

48 Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

49 Tuomas Virtanen. *Sound Source Separation in Monaural Music Signals*. PhD thesis, Tampere University of Technology, 2006.

50 Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, 2007.

51 Ron J. Weiss and Juan Pablo Bello. Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 123–128, Utrecht, The Netherlands, 2010.

52 John Woodruff and Bryan Pardo. Using pitch, amplitude modulation, and spatial cues for separation of harmonic instruments from stereo music recordings (article id 86369). *EURASIP Journal on Advances in Signal Processing*, 2007.

53 John Woodruff and Bryan Pardo. Active source estimation for improved source separation. Technical Report NWU-EECS-06-01, Electrical Engineering and Computer Science Department, Northwestern University, 2006.

54 John Woodruff, Bryan Pardo, and Roger B. Dannenberg. Remixing stereo music with score-informed source separation. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 314–319, 2006.

55 Jun Wu, Emmanuel Vincent, Stanislaw Andrzej Raczynski, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama. Multipitch estimation by joint modeling of harmonic and transient sounds. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 25–28, Prague, Czech Republic, 2011.

56 Shangming Yang and Zhang Yi. Convergence analysis of non-negative matrix factorization for BSS algorithm. *Neural Processing Letters*, 31:45–64, 2010.

Music Information Retrieval Meets Music Education

Christian Dittmar[1], Estefanía Cano[2], Jakob Abeßer[3], and Sascha Grollmisch[4]

1 **Semantic Music Technologies Group, Fraunhofer IDMT**
98693 Ilmenau, Germany
dmr@idmt.fraunhofer.de
2 cano@idmt.fraunhofer.de
3 abr@idmt.fraunhofer.de
4 goh@idmt.fraunhofer.de

Abstract

This paper addresses the use of Music Information Retrieval (MIR) techniques in music education and their integration in learning software. A general overview of systems that are either commercially available or in research stage is presented. Furthermore, three well-known MIR methods used in music learning systems and their state-of-the-art are described: music transcription, solo and accompaniment track creation, and generation of performance instructions. As a representative example of a music learning system developed within the MIR community, the Songs2See software is outlined. Finally, challenges and directions for future research are described.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems, I.5 Pattern Recognition

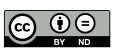
Keywords and phrases Music learning, music transcription, source separation, performance feedback

Digital Object Identifier 10.4230/DFU.Vol3.11041.95

1 Introduction

The rapid development of music technology in the past decades has dramatically changed the way people interact with music today. The way people enjoy and relate to music has changed due to the enormous flexibility given by digital music formats, the huge amount of available information, the numerous platforms for searching, sharing, and recommending music, and the powerful tools for mixing and editing audio.

Consequently, the potential of applying such technologies to music education was recognized. An automatic system that could potentially give instructions and feedback in terms of rhythm, pitch, intonation, expression, and other musical aspects could become a very powerful teaching and learning tool. However, in the early years between the 1980s and the early 2000s, automatic methods for pitch detection, music transcription, and sound separation among other methods, were still in very preliminary stages. Consequently, initial systems for music education, even though innovative and creative, had many restrictions and mainly relied on the possibilities offered by recording studios. In the late 1980s, play-along CDs became popular and offered a number of specially recorded tracks where the user could play with the provided accompaniment. Furthermore, instructional videos were recorded, which mainly featured famous musicians that offered some guidelines in terms of performance and practice. Later on, and mainly aiming for entertainment and not explicitly for music



Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 95–120

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

education, the video game community approached music with rhythm games that required the user to follow and repeat patterns of fingering gestures on special hardware controllers. The first commercial music rhythm game dates back to 1996 [35]. Even though these systems were not specifically created as educational tools, they were and still are particularly successful in creating interest in music performance and thus play an educational role.

Interactive applications with a more formal approach to music education have been created such as web services and software tools that guide students through different musical topics like music history or musical instruments. These systems mainly find use in music schools and universities as part of their class work and usually present a set of predefined lectures or practices that the students need to complete.

Recent developments in portable devices like smart-phones and tablets resulted in higher processing power, more powerful audio processing features, and more appealing visuals. As a result, the app market has had an immense growth and everyday more music-related applications are available for both the Android and iOS market. The MIR community has had its share in the development of pitch detection, audio recommendation, and audio identification algorithms necessary for such applications.

The usage of music technology in music education is an ongoing process: on the one hand it completely relies on the accomplishments of the scientific community; on the other hand, it is a process that requires a progressive change of mentality in a community where many processes and techniques still remain very traditional. The development of new music education systems faces many challenges: (1) Development of music technologies robust and efficient enough to be delivered to the final user. (2) Bridging the gap between two communities—music education and music technology—that have completely different environments and mentalities. (3) Design of appealing and entertaining systems capable of creating interest while developing real musical skills.

The remainder of this paper is organized as follows: Section 2 describes some relevant systems for music education, Section 3 presents three Music Information Retrieval (MIR) methods applied in music education applications, Section 4 describes Songs2See—a current music education system developed within the MIR community. Finally, Section 5 discusses future challenges faced by the music information retrieval community and Section 6 draws some conclusions.

2 Related Systems for Music Education

This section gives a general overview of music education systems that are either commercially available or representative of the state-of-the-art in the MIR community. In general, music education systems can be broadly classified in three categories: play-along CDs and instructional videos, video games, and software for music education.

2.1 Published Music Education Material

Starting in the 1980s, play-along CDs and instructional videos became popular as an alternative way to practice an instrument. Play-along CDs consist of specially recorded versions of popular musical pieces, where the user plays along to the recorded accompaniment. The main advantage of these systems is that users can practice with their own musical instrument: any progress is directly achieved by real instrumental practice. Furthermore, these systems allow users to get familiar with accompaniment parts—for example, piano or orchestra accompaniments—and as such, they became a popular tool for concert and contest preparation. On the other hand, the amount of available content is limited by the

particularly high production costs of such systems: in many cases, large ensembles and long recording sessions are needed for the production of one track. In this sense, play-along CDs are mainly available for very popular songs and for some representative concerts of the instrumental repertoire. Music Minus One[1], for example, offers a large catalog of play-along CDs for different instruments, ensembles, and genres. The Hal Leonard Corporation[2] has published a series of jazz play-alongs for different instruments, with compilations of music of different artists, jazz standards, and thematic editions. Another very popular series of play-alongs is the jazz series published by Jamey Aebersold[3] with a catalog of over a 100 items featuring different artists, playing techniques, and jazz standards.

Instructional videos came out as an educational tool where renowned musicians addressed particular topics—playing techniques, improvisation, warm-up exercises—and offered hints and instructions to help users to improve their skills and to achieve a certain goal. For theses cases, the popularity of a certain musician, as opposed to the popularity of a musical piece, was used as a marketing tool. The idea that you could play like famous musicians do, was very appealing. With time, the catalog of instructional videos grew both in size and diversity, featuring not only famous musicians, but also different playing techniques, learning methods, and the very famous self-teaching videos. The popular VHS tapes from the 1980s and 1990s where slowly replaced by digital formats like the VCD and DVD. Alfred Music Publishing[4], Berklee Press[5], Icons of Rock[6] and Homespun[7] all offer a series of instructional videos for different instruments and styles.

The main weakness of both play-along CDs and instructional videos is that there is no direct feedback for the user in terms of performance evaluation. Users have to completely rely on their own perception and assessment, which in case of beginners, can be a real challenge. However, these types of learning material have played a very important role as they offer an alternative way to practice at home, helping to keep the motivation for learning, and offering the flexibility of practicing on your own time, pace, and schedule.

2.2 Music Video Games

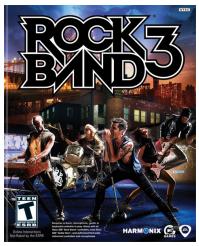
The 1990s was the decade where the development of music rhythm games[8] had a solid start, leading to the great popularity of music games in the next decade. The 1996 release and popularity gain of the game PaRappa the Rapper for Sony PlayStation 1 was an important propeller of the music game development[9]. Later examples of popular releases in the music video game community are Guitar Hero[10], Rock Band[11], and the karaoke game SingStar[12]. Guitar Hero has been

[1] Music Minus One: http://www.musicminusone.com
[2] Hal Leonard Corporation: http://www.halleonard.com/promo/promo.do?promotion=590001&subsiteid=1
[3] Jamey Aebersold: http://www.jazzbooks.com/jazz/category/AEBPLA
[4] Alfred Music Publishing: http://www.alfred.com/Browse/Formats/DVD.aspx
[5] Berklee Pree: http://www.berkleepress.com/catalog/product-type-browse?product_type_id=10
[6] Icons of Rock: http://www.livetojam.com/ltjstore/index.php5?app=ccp0&ns=splash
[7] Homespun: http://www.homespuntapes.com/home.html
[8] Type of music video games that challenge a player's sense of rhythm. They usually require the user to press a sequence of buttons shown on the screen.
[9] PaRappa the Rapper: http://www.gamestop.com/psp/games/parappa-the-rapper/65476
[10] Guitar Hero: http://www.guitarhero.com
[11] Rock Band: http://www.rockband.com
[12] Singstar: http://www.singstar.com

released for different video game consoles like Microsoft Xbox 360, Sony PS3, and also for Windows PCs. The series started as a pure rhythm guitar music game in which the user had to press the correct button at the right time, requiring rapid reflexes and good hand-to-eye coordination [23]. The controller resembles a real guitar but instead of strings and frets, it has five buttons and a strum bar.

(a) Guitar Hero

(b) Singstar



(c) Rock Band 3

(d) Rocksmith

Figure 1 Music Rhythm Games.

Rock Band 3 has been released for Microsoft Xbox 360, Nintendo Wii, Sony PS3, and Nintendo DS. It supports up to three singers with a three-part harmony recognition feature. It was released with a set of 83 songs and has full compatibility with all Rock Band peripherals as well as most Guitar Hero instruments.

One important characteristic of the above mentioned rhythm games is that, while being entertaining and successful in creating interest in music performance, they often fail to develop musical skills that can be directly transfered to real musical instruments as game controllers cannot really capture complexities and intricacies of musical instruments.

(a) Guitar Hero guitars

(b) Rock Band 3 drum set

(c) Rock Band 3 keyboard

(d) Rock Band 3 guitar

Figure 2 Music game controllers.

SingStar was released for Sony PlayStation 2 & 3. Like conventional karaoke games, it offers the possibility to sing along to the included songs with the lyrics shown synchronously. Additionally, the singing input is automatically rated by the game, which requires the original vocal tracks to be transcribed beforehand by the producers of the game.

The first commercial release in the video game community that allows users to play with real guitars is Rocksmith[13], released in the United States in September 2011 for Microsoft Xbox 360, Windows, and Sony PS3. The system allows users to connect their guitar output via USB interface. The user's performance is rated based on analysis of the audio signal. Like other music games, Rocksmith delivers a set of songs specifically edited for the game. As an additional feature, Rocksmith offers a series of mini-games with scales, finger dexterity,

[13] Rocksmith: `http://rocksmith.ubi.com/rocksmith/en-US/home/`

or chord exercises for developing playing skills. This represents a major leap in the music game community as no special hardware controllers are needed and, instead of following button sequences, the user actually plays a real guitar with the fret and string information provided in the game. Furthermore, the inclusion of additional exercise-games paves the way from mere gaming and entertainment to music education.
A common limitation of the music games mentioned above is that content is entirely limited to a set of songs delivered with each game. Even if a great effort is made to deliver popular songs appealing to a wide audience, personal tastes can never be completely covered. Furthermore, content is in general limited to pop and rock releases, ignoring the large amount of other musical genres and styles.

2.3 Music Education Software

In terms of commercial systems for music education, Music Delta[14], Smart Music[15] and GarageBand[16] present interactive alternatives to music learning. Music Delta is a web based system developed by Grieg Music Education comprising music curricula, content articles, and interactive tools. There are two versions available: (1) Music Delta Master, an online textbook which offers different performance stages and a special tool for composing and remixing. (2) Music Delta planet, specially designed for elementary school children where topics as music history and composers are presented in an entertaining way.
SmartMusic is a Windows and Mac software developed by MakeMusic especially for bands, orchestras, and vocals. Users can play their instruments to the computer microphone and receive immediate feedback from the software. One of the biggest strengths of the system is that teachers can assign tasks for the students to practice at home. The student's progress can be tracked and personal rating system can be generated. Currently, there are around 2000 musical pieces available for the software.
GarageBand is a software released by Apple for Mac and iPad. Among many other features, it provides the possibility to learn how to play piano and guitar with specially designed content, performance feedback, and appealing user interfaces. Users can play directly to the computer microphone or through USB connection.

With a slightly different approach, Songle[17] offers a web service for active music listening. Users can select a song from the list or register to include an audio file via URL. The system uses MIR techniques to analyze the audio file and then displays information regarding melody line, chords, structural segments, and beat grid. As errors are expected in the automatic analysis, users can edit, correct, and include missing information [21]. The main idea behind this system is to allow users to have a deeper understanding of music and enrich the listening experience.

2.4 Music-Related Mobile Apps

As mentioned in Section 1, the development of apps for smartphones and tablets has grown very rapidly in the last years. Many music-related applications are already on the market, some of them dealing with music learning and playing. Rock Prodigy[18] is a guitar playing

[14] Music Delta: `http://www.musicdelta.com`
[15] Smart Music: `http://www.smartmusic.com`
[16] Garage Band: `http://www.apple.com/ilife/garageband/`
[17] Songle: `http://songle.jp`
[18] Rock Prodigy: `http://www.rockprodigy.com`

app developed for the iPad, iPhone, and iPod Touch. Users can play their guitar directly to microphone and, based on a lesson plan and a rating system, receive performance feedback from the application. The lesson plan offers chords, rhythm, scales, technique, and theory exercises. There are also popular songs available for purchase that can be played within the app. Tonara[19] is an interactive sheet music iPad app where users can download and view music directly on their iPad. The app records input directly from the microphone and automatically detects the user's position in the score as the user plays. The system can be used with any musical instrument but currently only violin, piano, flute, and cello scores are available for purchase. Wild Chords[20] is a music game developed by Ovelin and designed to help beginners familiarize with guitar chords. It is available as an iPad app and uses appealing visuals and references to animals to help users identify the chords. The game records audio input directly from the microphone and no hardware controllers are needed.

2.5 Research Projects

In the past years a few research projects have dealt with the development of E-learning systems for music education. The IMUTUS [21] (Interactive Music Tuition System), the VEMUS [22](Virtual European Music School), and the i-Maestro [23] (Interactive Multimedia Environment for Technology Enhanced Music Education and Creative Collaborative Composition and Performance), were all European based projects partially funded by the European Commission that addressed music education from an interactive point of view. IMUTUS focused on the recorder with the goal developing a practice environment where students could perform and get immediate feedback from their renditions. VEMUS was proposed as a follow up project of IMUTUS and addressed the inclusion of further musical instruments and the development of tools for self-practicing, music teaching, and remote learning. i-Maestro focused on the violin family and besides offering enhanced and collaborative practice tools, the project also addresses gesture and posture analysis based on audio visual systems and sensors attached to the performer's body.

Music Plus One [44] is a system for musical accompaniment developed in the attempt to make computer accompaniments more aesthetic and perceptually pleasing. It was developed in the School of Informatics & Computing in Indiana University. The idea behind the system is that a computer-driven orchestra listens, learns, and follows the soloist's expression and timing. The system is composed of three main blocks: (1) Listen: based on a Hidden Markov Model, this stage performs real-time score matching by identifying note onsets. (2) Play: generates an audio output by phase vocoding a pre-existing audio recording (3) Predict: predicts future timing by using a Kalman filter-like model.

Antescofo [9] is both a score-following system and a language for musical composition developed by the Music Representation Research Group at IRCAM. It allows automatic recognition of the player's position and tempo in a musical score. Antescofo can be used in interactive accompaniment scenarios and as a practicing tool. It can also be used as a composition tool by synchronizing and programming electronic events and computer generated sounds with instrumental performances. It also serves as a research tool for tempo and performance analysis.

---

[19] Tonara: `http://tonara.com`
[20] Wild Chords: `http://www.wildchords.com`
[21] IMUTUS: `http://www.exodus.gr/imutus/index.htm`
[22] VEMUS: `http://www.tehne.ro/projects/vemus_virtual_music_school.html`
[23] i-maestro: `http://www.i-maestro.org`

Songs2See [8] is a software developed within a project that started in 2010 at the Fraunhofer Institute for Digital Media Technology (IDMT)[24]. The main goal was to apply state-of-the-art MIR techniques in the development of a music education and learning tool. It takes advantage of pitch extraction, music transcription, and sound separation technologies to allow the user to play with real musical instruments—guitar, bass, piano, saxophone, trumpet, flute and voice. The system returns immediate performance feedback based on a rating system, and gives the user flexibility in terms of the content that can be used with the application—the user can load audio files to create content for the game. In Section 4, a thorough description of Songs2See [25] is presented.

A slightly different approach is taken in the project KOPRA-M[26], that started in 2011 at the Fraunhofer Institute for Digital Media Technology (IDMT). This project is focused on measurement of competencies in music. For this matter, a systematic methodology and a proprietary software solution to assign and control music tasks is developed. The outcomes of this project are targeted to German secondary school students. An important aspect is the (semi-)automatic scoring procedure for evaluating different singing and rhythm tasks. By employing MIR methods, an attempt is made to model ratings given by human experts through regression methods.

3 MIR Methods for Music Learning Applications

3.1 Music Transcription

Music transcription refers to the process of automatically extracting parameters such as pitch, onset, duration, and loudness of the notes played by any instrument within a recorded piece of music. Furthermore, rhythmic and tonal content provided by the beat grid and the musical key, are also of importance when transferring note sequences into common music notation. We refer to these parameters as *score parameters* since they generally do not make assumptions on the particular instrument that is notated.

The automatic transcription of a music piece is a very demanding task since it embraces many different analysis steps such as instrument recognition, multiple fundamental frequency analysis, and rhythmic analysis [30]. Depending on the type of musical instrument to be transcribed, music transcription algorithms are often associated with melody transcription, polyphonic transcription, drum transcription, or bass transcription [11]. The challenges in automatically transcribing music pieces are diverse. First, music recordings usually consist of multiple sound sources overlapping constructively or destructively in time and frequency. Mutual dependencies between the different sources exist due to rhythm and tonality. Furthermore, the number of sound sources is in general unknown and not easily extracted and consequently, often needs to be given as a parameter to the algorithms. Second, all sound sources have very diverse spectral and temporal characteristics that strongly depend on the instrument type and the applied playing techniques (see also Section 3.1.1). Finally, different instruments can be associated with different functional groups such as the main melody or the bass line.

Transcribing the main melody is the most popular task due to the various applications in music education or karaoke systems. If multiple melodies are played simultaneously, different perceptual criteria need to be considered by the transcription algorithm in order to identify

[24] Fraunhofer IDMT: `http://www.idmt.fraunhofer.de/en.html`
[25] Songs2See: `http://www.songs2see.com`
[26] KOPRA-M: `http://www.idmt.fraunhofer.de/de/projekte/laufende_projekte/KOPRA-M.html`

the main melody. A selection of existing transcription algorithms are thoroughly described in [20], [4], and [3]. In general, state-of-the art automatic music transcription algorithms consist of the following parts:

- **Time-frequency representation:** In order to separately analyze frequency components of different instruments, a suitable time-frequency representation needs to be computed. Commonly used techniques are the Short-time Fourier Transform (STFT), the Multi-Resolution FFT (MR-FFT) [12], the Constant-Q Transform [5], or the resonator time frequency image (RTFI) [56].
- **Spectral decomposition:** Often based on harmonic templates, techniques such as Non-Negative Matrix Factorization (NMF) [33] or Probabilistic Latent Component Analysis (PLCA) [48] are used to decompose the time-frequency representation into the contributions of different (harmonic) instruments. Spectral decomposition yields one or multiple fundamental frequency estimates for each time frame.
- **Onset detection & note tracking:** Probabilistic models such as Hidden-Markov Models (HMM) [41] are applied to model the temporal progression of notes and to estimate their onset and offset time. Based on the frame-wise fundamental frequency estimates, the pitch can be extracted for each note event.

The Music Information Retrieval Evaluation eXchange (MIREX[27]) contest offers several transcription-related tasks such as "Audio Melody Extraction", "Multiple Fundamental Frequency Estimation & Tracking", and "Audio Onset Detection". In this annual contest, various algorithms based on signal processing techniques are evaluated and compared.

In the context of music education, automatic music transcription is an indispensable tool for the automatic generation of music exercises from arbitrary recordings. Music transcription applications allow to detect playing errors in real-time. Thus, the musical performance can be evaluated immediately. By using these applications, music students are not restricted to take lessons in the environment of a music school anymore. Instead, they can use automatic learning tools always and everywhere, which increases their motivation and enhances their musical experience.

3.1.1 Extraction of Instrument-Specific Parameters

In contrast to the *score parameters* discussed in the previous section, there are other parameters that describe performance aspects on a specific instrument. These *instrument-specific* parameters provide cues about the applied playing techniques such as finger-style guitar play or slap-style bass guitar play. They can also describe different techniques such as vibrato or string bending used by the musician as expressive gestures during the performance. These techniques alter the fundamental frequency of a note in a characteristic way and can be parametrized with different levels of granularity, depending on the context of application [1]. Some studies focus solely on estimating *instrument-specific* parameters from audio recordings whereas other studies use these parameters to improve music synthesis algorithms. In this section, four selected studies are briefly discussed that focus on the clarinet, the classical guitar, and the bass guitar.

Sterling et al. [49] presented a physical modelling synthesis algorithm that incorporates two typical performance gestures for clarinet synthesis—tonguing and the pitch bend. Pitch bending allows the musician to alter the fundamental frequency within a range of about a

[27] http://www.music-ir.org/mirex/wiki/2011:Main_Page

semitone. Tonguing relates to the note articulation and controls the onset and offsets while maintaining a constant blowing pressure.

In [38], Özaslan and Arcus focused on two expression styles performed on the classical guitar—the legato technique, which corresponds to the hammer-on and pull-off techniques, and the glissando technique, which corresponds to the slide technique. Both techniques result in an ascending or descending pitch after the note is initially plucked. The authors use a high frequency content (HFC) measure to detect the pluck onset with its percussive characteristics. Between note onsets, the YIN fundamental frequency estimation algorithm [10] was used to characterize the pitch progression. The note release part was segmented into five segments to analyze whether a continuous or an abrupt change of the fundamental frequency appears.

Laurson et al. [32] presented a method for synthesizing the use of rasgueado technique on the classical guitar, which is a rhythmically complex strumming technique primarily used in flamenco music. The technique is characterized by fast consecutive note plucks with the finger nails of all five fingers of the plucking hand and an upwards and downwards movement of the plucking hand. The authors extract signal characteristics—timing and amplitude of individual note attacks—from real-world recordings and use it to re-synthesize the recording.

Abeßer et al. [2] presented a feature-based approach for automatically estimating the plucking style and expression style from isolated bass guitar notes. The authors used a taxonomy of ten playing techniques and described several audio features that capture specific characteristics of the different techniques. The parameter estimation is interpreted as a classification problem.

3.1.2 Spatial Transcription Using Different Modalities

In addition to the score parameters and instrument-specific parameters discussed in the previous sections, two more aspects need to be considered in order to get a better semantic description of a performed instrument track. Firstly, when music is performed on a string instrument such as the bass or the guitar, a given set of notes can usually be played on different positions on the instrument neck. Due to the tuning of the strings and the available number of frets, the assignment between note pitch and *fretboard position*—fret number and string number—is ambiguous. Second, different fingers of the playing hand can be used in order to play a note within a fixed fretboard position. Consequently, various *fingerings* are possible to play the same set of notes. This fact holds true also for other instruments such as the piano and the trumpet. The process of finding the optimal fingering is discussed in Section 3.3.2.

In order to chose suitable fretboard positions and fingerings, musicians usually try to minimize the amount of physical strain that is necessary to play an instrument. For string instruments, this strain is caused by finger stretching and hand movement across the instrument neck. We refer to *spatial transcription* as the process of estimating the applied fretboard positions. In order to automatically estimate the fretboard positions and the fingering from a given musical recording, the sole focus on the audio analysis is often not sufficient. This is mainly due to the fact that the change of fingering barely alters the sonic properties of the recorded signal. In the following sections, we discuss different approaches that include methods from computer vision as a multi-modal extension of audio analysis.

3.1.2.1 Audio-visual Approaches

Already in 2003, Smaragdis and Casey [47] proposed the use of audio-visual data for the extraction of independent objects from multi-modal scenes. They utilized early fusion by means of concatenating audio spectra and corresponding image frames and using Independent Component Analysis (ICA). Although merely a proposal, their paper included an early practical example: onset detection of piano notes by analyzing both a recorded video of the played keyboard and the corresponding audio signal.

Hybryk and Kim [25] proposed a combined audio and video analysis approach to estimate the fretboard position of chords that were played on an acoustic guitar. First, the authors aim at identifying all played chords in terms of their "chord style", i.e., their root note and mode such as E minor. The Specmurt algorithm [46] is used to analyze the spectral distribution of multiple notes with their corresponding harmonics. The outcome is a set of fundamental frequencies that can be associated to different note pitches. The amplitude weights of the harmonic components are optimized for different pitch values. Based on the computed "chord style" (e.g., E minor), the "chord voicing" is estimated by tracking the spatial position of the gripping hand. The chord voicing describes the way the chords are played on the fretboard.

Paleari et. al. [39] presented a method for multimodal music transcription of acoustic guitars. The performing musicians were recorded using two microphones and a digital video camera. In addition to audio analysis, the visual modality was used to track the hand of the guitar player over time and to estimate the fretboard position. The fretboard position is initially detected by analyzing the video signal and then spatially tracked over time. Furthermore, the system detects the position of the playing hand and fuses the information from audio and video analysis to estimate the fretboard position.

3.1.2.2 Visual Approaches

Other approaches solely use computer vision techniques for spatial transcription. Burns and Wanderley [6] presented an algorithm for real-time finger tracking. They use cameras attached to the guitar in order to get video recordings of the playing hand on the instrument neck. Kerdvibulvech and Saito [28] use a stereo camera setup to record a guitar player. Their system for finger tracking requires the musician to use *colored fingertips*. The main disadvantage of these approaches is that both the attached cameras as well as the colored fingertips are unnatural for the guitar player. Therefore,this may influence the user's expressive gestures and playing style.

3.1.2.3 Enhanced Instruments

A different approach is followed when using *enhanced music instruments* that comprise additional sensors and controllers to directly measure the desired parameters instead of estimating them from an audio or video signal. The major disadvantage of enhanced instruments is that despite of their high accuracy in estimating performance and spatial parameters, they are obtrusive to the musicians and may affect their performance on the instrument [25].

Music game controllers as depicted in Figure 2 were introduced as parts of music games such as Guitar Hero or Rockband. These controllers imitate real instruments in shape and functions and are usually purchased in combination with the corresponding game. However, the controllers often simplify the original musical instruments. The Guitar Hero controller, for instance, reduces the number of available frets on the instrument track from 22 to 4 and

furthermore encodes each fret with a different color. The player does not pluck strings but instead presses colored buttons. These simplifications reduce the complexity of performance instructions within the music games and guarantee faster learning success for beginners. The main disadvantage of such controllers is that even though they have a similar instrument shape, their way of playing differs strongly from real instruments. Thus, learning to use the controllers does not necessarily help when learning to play a real instrument.

Hexaphonic pickups allow guitar players and bass guitar players to use their instruments as MIDI input devices, a feature otherwise only available to keyboard players using MIDI keyboards. Since each instrument string is captured individually without any spectral overlap or additional sound sources, a fast and robust pitch detection with very low latency and very high accuracy can be realized using the individual pickup signals as input. This transcription process is usually implemented in an additional hardware device. This way, hexaphonic pickup signals can be converted into MIDI signals nearly in real-time. These MIDI signals allow the musician to intuitively play and control sequencer software, samplers, or virtual instruments in real time.

3.2 Solo & Accompaniment Track Creation

The idea that users can take any recording of their choice—no matter how, where, and when it was created—and obtain solo and accompaniment tracks to play along with, represents a very appealing and flexible approach for practicing music. Whether it be playing along to the Berlin Philharmonic or to the Count Basie Orchestra, all would be possible. Besides being a powerful practicing aid, solo and accompaniment tracks can also be used for performance and musicological studies. We consider sound source separation as a common ground for solo and accompaniment track creation algorithms and further describe it in the next section.

3.2.1 Source Separation

In the context of solo and accompaniment track creation, some separation systems have specifically focused on singing voice extraction from polyphonic audio—the solo instrument is always assumed to be the singing voice. In [34], a system based on classification of vocal/non-vocal sections of the audio file, followed by a pitch detection stage and grouping of the time frequency tiles was proposed. In [45], voice extraction is achieved by main melody transcription and sinusoidal modeling. A system based on pitch detection and non-negative matrix factorization (NMF) is proposed in [52]. Others have focused on the separation of harmonic from percussive components of an audio track [37], [19]. Similarly, a system is proposed in [7] to specifically address the extraction of saxophone parts in classical saxophone recordings and in [16], a score-guided system for left and right hand separation in piano recordings is proposed. More general algorithms have also been proposed for main melody separation regardless of the instrument used: Durrieu [14] proposes a source/filter approach with a two-stage parameter estimation and Wiener filtering based separation. In [31], Lagrange proposes a main melody extraction system based on a graph partitioning strategy—normalized cuts, sinusoidal modeling and computational auditory scene analysis (CASA).

In [8], a system for solo and accompaniment track creation is presented. This algorithm is included in the Songs2See application (see Section 4). The system is composed of five building blocks shown in the diagram in Figure 3. It was designed with the goal of taking audio files from commercial recordings and by means of a pitch detection algorithm and a sound separation scheme, identify the predominant melody in the track, extract the main

melody and deliver two independent tracks for the accompaniment and solo instrument. The different processing blocks are briefly explained:

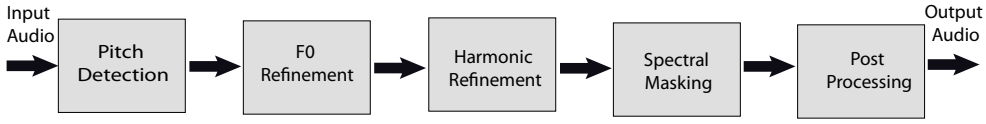
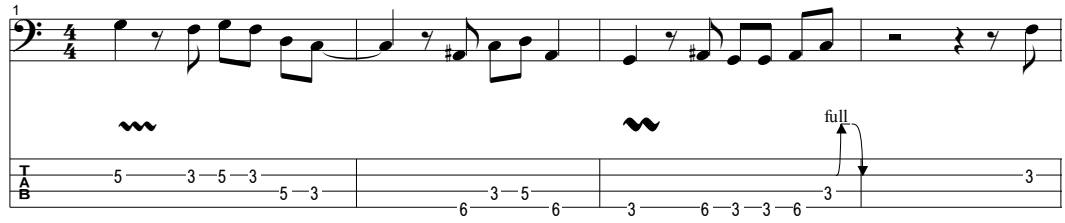
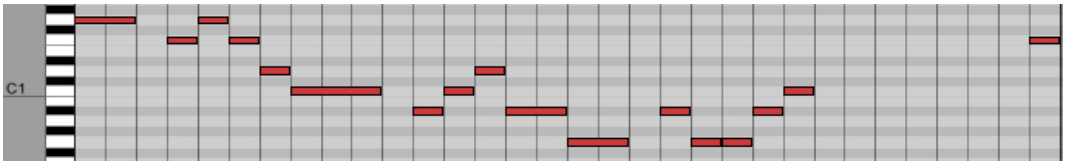


Figure 3 Block diagram of the solo and accompaniment track creation algorithm.

- **Pitch Detection:** The pitch detection algorithm proposed in [13] is used, which uses a multi-resolution FFT as a front end. After an initial peak detection stage, pitch candidates are obtained based on a pair-wise evaluation of spectral peaks. Tones are formed based on information of past analysis frames, gathered to analyze long term spectral envelopes, magnitude and pitch information. The main voice is obtained from the pitch candidates using a salience measure.
- **F0 Refinement:** To further improve F0 estimations, a refinement stage is proposed where the magnitude spectrogram is interpolated in a narrow band around each initial F0 value and its constituent harmonics. To obtain a more realistic estimate of the harmonic series, an inharmonicity measure is introduced where harmonic components are not expected to be exact integer multiples of the fundamental frequency but may slightly deviate from the theoretic values.
- **Harmonic Refinement:** The underlying principle at this stage is that the higher the harmonic number of a partial, the more its location will deviate from the calculated harmonic location, i.e., multiple integer of the fundamental frequency. Three main aspects are considered here: (1) Each harmonic component is allowed to have an independent deviation from the calculated harmonic location. (2) Each partial is allowed to deviate from its harmonic location a maximum of one quarter tone. (3) Acoustic differences of string and wind instruments are considered.
- **Spectral Filtering:** After the complete harmonic series has been estimated, an initial binary mask is created where each time-frequency tile is defined either as part of the solo instrument or the accompaniment. To compensate for spectral leakage in the time frequency transform, a tolerance band is defined where not only the specific time frequency tile found in the harmonic refinement stage is filtered out but also the tiles within a band centered at the estimated location.
- **Post Processing:** A final refinement stage is implemented where the different tones are independently processed to remove attacks and possible artifacts caused by inaccurate classification of the time-frequency tiles. Two cases are considered: (1) Due to the particular characteristics of the pitch detection algorithm, a few processing frames are necessary before a valid F0 value is detected. This sometimes causes the pitch detection algorithm to miss the attack frames that belong to each tone. To compensate for the inherent delay in the pitch detection algorithm, a region of 70 ms before the start of each tone is searched for harmonic components that correlate with the harmonic structure of each tone. The binary mask is modified accordingly to include the attack frames found for each tone. (2) Percussion hits are often mistakenly detected as being part of a tone. To reduce the perceptual impact of these inaccuracies, a final analysis of the tone is performed where the harmonic series is analyzed as a whole and transients occurring



(a) Score, tablature



(b) Piano roll

Figure 4 Score, tablature, and piano-roll representation of a bass-line.

in several harmonic components simultaneously are detected. For these time-frequency tiles, the spectral mask is weighted and is no longer binary. Finally, the complex valued spectrogram is masked and independent solo and accompaniment tracks are re-synthesized by means of an inverse short term Fourier transform.

3.3 Performance Instructions

3.3.1 Music Representations

In this section, three different symbolic music representations are compared. First, we briefly review the *score* and the *tablature* representations since they are the two most popular written representations of a music pieces. Afterwards, we discuss the *piano-roll* representation, which is often used in music production software, music education applications, as well as in music games. In the three representations, each note is described as a temporal event that is characterized by a set of distinct parameters such as note pitch or duration. As an example, a bass line is illustrated as score, tablature, and piano-roll representation in Figure 4.

The score notation is the oldest and most popular written representation of music. It offers a unified and well-established vocabulary for musicians to notate music pieces for different instruments. Furthermore, the score notation provides a compact visualization of the rhythmic, harmonic, and structural properties of a music piece.

The tablature representation, on the other hand, is specialized on the geometry of fretted string instruments such as the guitar or the bass guitar. Each note is visualized according to its fretboard position, i.e., the applied string number and fret number. Due to the instrument construction and tuning of most string instruments, notes with the same pitch can be played in different fretboard positions. The choice of the fretboard position usually depends on the stylistic preferences of the musician and the style of music. Tablatures often include additional performance instructions that indicate the playing techniques for each note. These techniques range from frequency modulation techniques such as vibrato or bending to plucking techniques such as muted play or slap style for the bass guitar. The main advantage of the tablature representation is that it resolves the ambiguity between note pitch and fretboard position. This benefit comes along with several problems: (1) Tablatures are hard to read for musicians who play other instruments such as the piano, trumpet, or saxophone.

(2) Tablatures often do not contain any information about the rhythmic properties of notes. Different note lengths are sometimes encoded in different distances between the characters but this method is often ambiguous. (3) Tablatures, which are nowadays easily accessible from the Internet, are often incomplete or erroneous because they were written by semi-professional musicians. Without the help of a teacher, music students often cannot identify the erroneous parts. Instead, the students might adopt inherent mistakes without even being aware of it.

Finally, the piano-roll representation is often used in music sequencer programs where each note is represented as a rectangle. The rectangle's length encodes the note duration and its horizontal and vertical coordinates encode the note onset and pitch, respectively.

All three representations discussed in this section fail to provide information on micro-tonality or micro-timing of a music piece. These aspects can usually be neglected in basic music education where the main focus of study is more on learning to play melodies than to imitate a particular performance style. Once the student reaches a certain skill level, the precise imitation of a given performance and artistic shaping becomes more important. Here, even slight differences in terms of micro-tonality (intonation) or micro-timing can be of high importance.

3.3.2 Automatic Generation of Fingering Instructions

Fingering instructions indicate which finger or set of fingers of the playing hand are to be used in order to play a particular note on an instrument. Both score notation and tablatures can provide this information by means of a number printed on top of each note. This number encodes the index of the finger that is to be used. However, most often, this information is not given. Since multiple fingerings are usually possible, finding the most convenient fingering is a challenging task. In general, fingering instructions can be generated manually or automatically. Trained music experts can derive optimal fingerings manually based on their experience. Even though this approach leads to proper performance instructions, it is time consuming and inapplicable for a fast analysis of a large number of songs. Furthermore, this manual process clearly stands in contrast to the idea of a complete automatic music transcription system. Therefore, automatic algorithms for fingering generation were developed based on the same criteria that musical experts apply.

In terms of applicable fingerings, musical instruments can be categorized into three different types. Instruments of the first type, such as the piano, have an 1-to-N relationship between note pitch values and possible fingerings. Each note pitch is generated by one distinct key but each key can be pressed by different fingers. Instruments of the second type, such as the saxophone or the trumpet, have an N-to-1 relationship between pitch and fingering. These instruments can produce each pitch with possibly different fingerings but each fingering has a unique finger assignment. For instance, each key on the saxophone is associated to one finger but the same note pitch can be played by using different key combinations. These combinations require a different amount of physical strain both for gripping and blowing the instrument. Similar to the fretboard positions for string instruments discussed in Section 3.1.2, the choice of the fingering depends on the stylistic preferences, performance level, and the musical context within a music piece. The most complex case is the third type of instruments with an N-to-N relationship such as the guitar or the bass guitar. On these instruments, each note can be played at different positions on the instrument neck as well as using different fingers.

Algorithms for automatic fingering generation have to be tailored towards the specific properties of musical instruments, including geometry, tuning, and pitch range. Furthermore, these algorithms have to be scalable to music pieces of different lengths. Usually, a cost value

is assigned to each possible fingering in order to automatically derive the optimal fingering.

The process of manually selecting the optimal fingering is influenced by different factors that are part of an underlying cognitive process of the musician [40]. In this study, the authors focused on piano fingerings but discussed several factors that can be applied to other instruments:

- **Biomechanical criteria:** These criteria relate mostly to physical strain, i.e., the necessary effort to play a note on the instrument.
- **Cognitive criteria:** These criteria are often related to the given musical context such as the rhythmic structure of a piece. As an example, strong accents are usually played with stronger fingers.
- **Conventional criteria:** One of these criteria, for example, indicates that musicians prefer using fingering patterns that they already learned over new patterns.
- **Additional criteria:** These criteria comprise musical style, timbre, and intonation and describe how these factors are affected by different fingerings.

Furthermore, the skill level of the musician strongly influences the choice of fingerings. Algorithms for automatic fingering generation need to include these criteria for generating usable results.

Most methods found in the literature focus on the guitar and the piano likely due to their high popularity. For polyphonic guitar fingerings, the presented methods usually distinguish between two different types of costs as in [27], [42], and [43]. Horizontal costs describe the difficulty of changing between different fretboard positions and vertical costs describe the complexity of a chord shape at a fixed position. Kasimi et al. [27] use the Viterbi Algorithm to determine the optimal fingering for a given note sequence. In contrast, Tuohy and Potter [50] follow a two-step approach. First, they use a genetic algorithm to generate a tablature representation. Then, they apply a neural network that was trained based on expert knowledge to derive the optimal fingering. Radisavljevic et al. [43] introduce the "path difference learning" approach, which allows to adapt the cost factors to the individual needs and abilities of the musician.

Presented methods for piano fingerings focus usually on short monophonic melody phrases. Hart et al. [24] assigned cost values to different transitions between white and black keys on the piano. In [40], Parncutt et al. empirically found that pianist in average, read eight notes ahead when playing monophonic melodies. The authors assigned cost values based on the size of intervals and used a set of twelve rules to derive an optimal fingering. Yonebayashi et al. [54] use a Hidden Markov Model to model different fingering positions and apply the Viterbi Algorithm to derive the final fingering.

Similar approaches to obtain optimal fingerings were discussed for the flute in [18] and for the trumpet in [26]. A detailed comparison of the presented methods can be found in [53].

4 Songs2See

Songs2See[28] is an application software for music learning, practice, and gaming which integrates various state-of-the-art MIR algorithms. In developing this system, the following requirements were taken into consideration:

1. The system should allow the use of real musical instruments or the voice without requiring special game controllers.

[28] http://www.songs2see.com

2. Users should be able to create their own musical exercise-content by loading audio files and using the analysis features available in the software.
3. The system should provide the entertainment and engagement of music video games while offering appropriate methods to develop musical skills.

4.1 System Architecture

Songs2See consists of two main applications: the Songs2See Editor, used to create content for the game, and the Songs2See Game, used at practice time. Figure 5 shows a block diagram of the system.

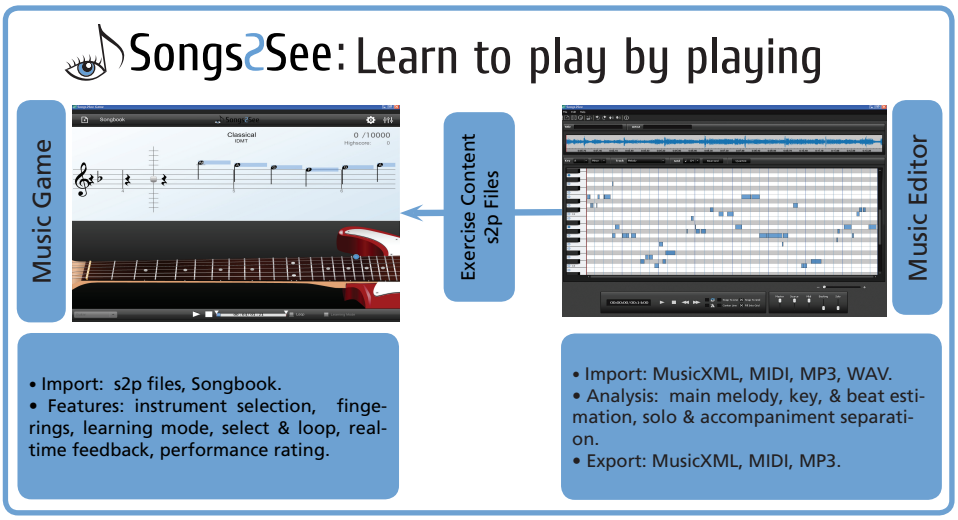


Figure 5 Songs2See block diagram.

The Songs2See Game is a platform-independent application based on Adobe Flash. The Songs2See Editor is a stand-alone application currently available for Windows PCs. The only additional hardware needed for Songs2See are speakers or headphones and a computer microphone to capture the performances. The standard work-flow in Songs2See is as follows: (1) Choose or create the content to be played: either select a track from the delivered songbook or load an audio file to the Songs2See Editor and use the analysis tools to create content for the game. The Editor sessions can be exported for the Game as .s2p files. (2) Load file in the Songs2See Game. (3) Select the desired instrument from the drop-down menu. (4) Run the Game and start playing. Besides the options already outlined, both the Songs2See Editor and Game offer several processing and performance options that will be further explained in the next sections.

4.2 Songs2See Game

The Songs2See Game is an application where users can practice a selected musical piece on their own musical instrument. There are several features in the Songs2See Game:

- **Score-like display of the main melody:** The Songs2See Game View, shown in Figure 6, combines elements both from standard piano roll views and score notation. The main goal was to include as many musical elements as possible without requiring the user to be able to read music beforehand. The length of the note is displayed both by using

music notation symbols—sixteenth notes, eight notes, quarter notes, half notes, whole notes, triplets, and their corresponding rests—and by displaying colored bars of different lengths behind each symbol. Note pitch is displayed both by placing the note objects in the correct position on the staff, and by writing the note names inside the note heads. The clef and key signature are also displayed on this view.

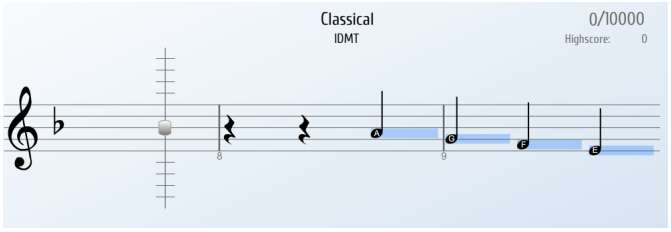
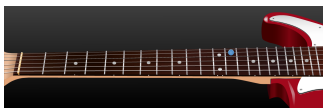


Figure 6 Game View.

- **Different instrument support:** The Songs2See Game currently supports bass, guitar, piano, saxophone, trumpet, flute, and singing voice. The user can select any of these instruments from the drop-down menu and an image of the instrument will be shown in the Instrument View. Figure 7 shows the different options for the instrument selection.

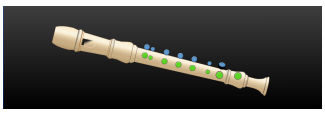


(a) Guitar

(b) Piano

(c) Trumpet

(d) Saxophone



(e) Flute

(f) Voice

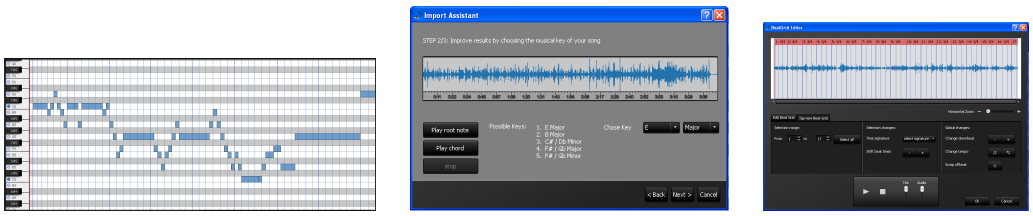
Figure 7 Instruments supported in Songs2See.

- **Instrument-specific fingering generation and display:** Every time the user loads a musical piece into the Game, the system automatically generates a fingering animation that describes the most common fingering sequence for the loaded melody. The fingering generation algorithm is instrument-specific and combines several criteria discussed in Sect. 3.3.2 for fingering selection. For instance, in the case of the trumpet, the blowing requirements are also displayed (see Figure 7c). For all instruments, the fingering for the current note in the note sequence is displayed in blue and the next note is displayed in green. This allows the user to prepare in advance for the next note in the melody sequence. In the event that the user wants to play a song on a musical instrument whose register does not allow to play all the notes—some notes might be too high or too low to be played in that particular instrument—a red sign will be displayed over the instrument to warn the user about register problems (see Figure 7d).
- **Real-time performance feedback:** Based on a real-time pitch detection algorithm [22], the user's performance is rated based on the pitch information extracted from the original audio track. When the user hits the note, the note will be painted in green and the user will score points with a maximum of 10000 points per song. When the user plays a wrong note, the error will be displayed and a reference for correction will be given.

- **Selection and looping option:** The user can select particularly difficult segments within the song and practice them separately. There is also a looping option where the selected segment will be played repeatedly.
- **Learning mode:** This option is meant to help users familiarize with the fingerings and performance requirements of a new song. When the Learning Mode is selected, the Game will be halted on each note until the user plays it correctly. This will give the user enough time to check finger positions, pitch, and all other performance details needed in the piece.
- **Songbook and loading options:** The user has two possibilities in terms of loading content into the Game: (1) The Songs2See Game is delivered with a set of songs that can be accessed through the Songbook icon. (2) Users can create their own content using the Songs2See Editor and exporting the session as an s2p file—the proprietary Songs2See format.
- **Other options:** Through the Options Menu, users can access several performance and set-up options: adjust delay between audio and visual to perfectly synchronize the Game with the performance, select the microphone input and adjust the gain, choose language, show or hide note names, enable or disable left hand mode for left-handed guitar players. In the Audio Menu, users can adjust the playback level as well as the balance between the accompaniment and solo tracks.

4.3 Songs2See Editor

The Songs2See Editor is a software component that allows users to create exercise content for the game. Additionally, it offers many general features that can potentially be used for many other applications outside Songs2See . The following options are available:



(a) Piano Roll View **(b)** Key Detection Assistant **(c)** Beat Grid Editor

Figure 8 Songs2See Editor: Piano Roll View, Key Assistant, and Beat Grid Editor.

- **Import options:** One of the most powerful features of the Songs2See Editor is that it allows the user to create material for the Game starting from different types and formats of audio material: WAV, MP3, MIDI, and MusicXML are supported. These four import possibilities make it possible to combine the use of the Songs2See Editor with other powerful processing tools as score-writing and sequencer software.
- **Main melody extraction:** Every audio file imported into the Songs2See Editor is automatically analyzed to extract the main melody. The employed algorithm [13] detects the most salient voice in the track regardless of the instrument played or the type of music. Results are displayed in the form of note objects in a Piano Roll View (see Figure 8a). As errors are expected, the user is allowed to create new notes, delete or merge existing notes, and adjust the length or the pitch of existing note objects. Audible feedback is

provided every time the user creates a new note or adjusts the pitch of an existing one. These options facilitate the process of correcting the melody extraction results.

- **Key detection:** The audio material imported into the Editor is also analyzed to extract key information. By using the Import Wizard (see Figure 8b), the user is presented with the five most probable keys and has the option to play the corresponding chords or root notes to help decision making. The key of the song can be changed at any time. The piano keys in the Piano Roll View that correspond to the notes in the selected key will be marked to guide the user through the editing process. The selected key is also important in terms of notation as this will be the key used when using the MusicXML export options.
- **Tempo and beat-grid extraction:** Tempo and beat extraction as presented in [11] are also performed when an audio file is loaded into the Editor. The automatic beat extraction currently only supports 4/4 time signatures. Considering possible extraction errors and the large amount of musical pieces written in other time signatures, the Songs2See Editor offers a processing tool called the Beat Grid Editor (see Figure 8c), where every beat can be independently modified and shifted. The downbeats can be modified and quick options for doubling or halving the tempo are available. A tapping option is also available, where the user can tap the beats of the song or sections of it.
- **Solo and accompaniment track creation:** The sound separation algorithm described in Section 3.2.1 is used to create independent solo and accompaniment tracks. The results from the main melody extraction are directly used in the separation stage. Therefore, the quality of the separation results is directly dependent on the quality of the extracted melody. The algorithm has been optimized to allow users to correct information in the melody extraction and receive immediate separation results based on the included changes.
- **Export options:** Every Songs2See Editor session can be exported for the Game as an s2p file—the proprietary format of Songs2See . All the necessary information for playback and performance is contained within this file. Furthermore, the intermediate results in the processing chain can also be exported to be used in other applications. For example, the solo and accompaniment tracks created can be exported as MP3 and then be used as learning and practicing material outside the Songs2See environment. The results from the main melody extraction can be exported as MusicXML or MIDI files and be used with score-writing or sequencer software.

5 Future Challenges

As described in the preceding sections, the state-of-the-art of MIR techniques for music education is already quite advanced. However, there are still numerous challenges present that motivate further research into specific directions. Some of them will be discussed in the following sections.

5.1 Polyphonic Music

Despite years of research, the automatic transcription of polyphonic music is still considered the holy grail in the MIR community. Many algorithms have been proposed that address polyphonic music played on monotimbral instruments, such as piano or guitar. The software

Celemony Melodyne[29] incorporates polyphonic transcription and sound separation of such data and is successfully used in music recording studios. During the last years, a multitude of novel signal processing front-ends for multi-pitch estimation have been proposed. Examples are Specmurt analysis [46] and systems as the ones presented in [29],[55],[4]. However, these methods only reveal pitch candidates and do not explicitly consider interference of un-pitched sounds such as drums. The majority of the proposed methods exhibit only straight forward post-processing based on empirical thresholds. One promising research direction is the usage of data-driven post-processing methods to train probabilistic models—such as HMMs. These methods rely on huge amounts of manually annotated audio material which can then be used to denoise the extracted multi-pitch candidates. These kinds of data can be derived from semi-automatic alignment of existing MIDI files to real-world recordings [17].

5.2 Sound Separation

Even though good results can be achieved in experiments under controlled or restricted conditions, most sound separation algorithms still lack robustness in real-world scenarios. A general solution, capable of identifying independent sound sources regardless of their timbre, salience, or recording conditions, has not been found.

With the recent development of perceptual measures for quality assessment [15], the sound separation community made an important step towards the development of structured and meaningful evaluation schemes. However, further research has to be conducted in terms of perceptual quality of extracted sources in order to better characterize algorithm artifacts and source interferences in terms of their perceptual impact. Separation evaluation campaigns as the Data Analysis Competition 05 have also been conducted since 2005[30]. The first separation campaign [51] specifically addressing Stereo Audio Source Separation (SASSEC) was conducted in 2007[31] and the SISEC (Signal Separation Evaluation Campaign) also included an audio separation task in 2011. The LVA/ICA[32] 2012 (International Conference on Latent Variable Analysis and Signal Separation), included a special session on real-world constraints and opportunities in audio source separation. These efforts represent an important step to push current state-of-the-art both in theoretical and practical aspects.

5.3 Improved Parametrization/Transcription

An improved parametrization of music recordings will be beneficial for various applications ranging from sound synthesis, music transcription, as well as music performance analysis. Each musical instrument has a unique set of playing techniques and gestures that constitute the "expressive vocabulary" of a performing musician. These gestures result in various sound events of different timbral characteristics. A detailed parametrization needs to be based on sophisticated models of sound production for each particular instrument. The applicability of these models can be evaluated by applying sound synthesis algorithm that are based on these models in order to re-synthesize the parametrized recordings. The main evaluation question remains: To what extend does a given model re-synthesize detected notes from a given recording while at the same time capturing the most important timbral characteristics of an instrument?

[29] Celemony Melodyne Editor: `http://www.celemony.com`
[30] Data Analysis Competition 05: `http://mlsp2005.conwiz.dk/index.php@id=30.html`
[31] SASSEC 07: `http://www.irisa.fr/metiss/SASSEC07/`
[32] LVA/ICA 2012: `http://events.ortra.com/lva/`

5.4 Non-Western Music

MIR methods have mainly addressed the characteristics of Western music genres. This is mainly due to the fact that state-of-the-art technologies were not robust enough to propose generalized solutions and constraints had to be placed in most methods to consistently address a particular problem. However, both due to the advances in the MIR community and to the spread of MIR research to Asian, African, and Latin American countries, a very strong interest in addressing other musics has emerged. New methods have to be developed to properly address the often complicated and intricate rhythmic and harmonic characteristics of these music styles. Collaboration between research facilities and experts from the different communities such as MIR and musicology is crucial to overcome current limitations.

5.5 Improvement of Computational Efficiency

One important obstacle for the inclusion of many MIR algorithms in commercial products is computational cost. Processing and time requirements are, in many cases, still very demanding, and even when they grant performance robustness, they also prevent the algorithms from being included in commercial applications. As in any research process, initial stages are always result-oriented. However, an effort has to be made to streamline algorithms to allow robust performance under standard computational capacities and facilitate real-time applications. As an example, under practical considerations, it is often sufficient to replace constant-Q spectra by conventional spectra obtained by Fast Fourier Transform subsequently resampled to a logarithmic frequency axis [36]. Although problematic from a signal theoretic point of view, this computationally efficient approach is often successfully used in music transcription methods.

5.6 Multi-User Applications

An important development direction for current music video games is the inclusion of multi-user applications that not only bring the entertainment and information contained within the game, but also further competition, engagement, and immersion from the interaction with other users. In music in particular, interaction with others is an expected scenario. Musicians rarely play alone and in most cases, they have to learn to interact and communicate with other musicians in order to produce an artistic ensemble performance. However, for multi-user applications to be feasible, algorithm efficiency has to be improved, real-time conditions have to be met, and latency and algorithmic delays reduced to the minimum. Furthermore, in the case of music learning applications, new feedback, rating and instruction approaches have to developed in order to properly assess the interaction and interplay with regard to intonation and timing.

5.7 Music Technology & Music Education

The inclusion of music technologies in both formal and informal music education is still fairly new. However, new generations grow up and live submerged in a digital era where possibilities are endless. This poses an important challenge to the music education community, as in order to reach the new generations, education methods have to evolve correspondingly. Nonetheless, changing mentalities and opening minds to new approaches is never an easy process and even less in a community as traditional as the music education community. This necessarily implies that music technology and music education have to work together to

reach a common goal: develop systems for music education that can be flexible, appealing, and suitable for developing real musical skills.

6 Conclusions

A general overview of the use of MIR technologies in music learning applications has been presented. Both the evolution of the community over time and its current state-of-the-art suggest that music education will be rapidly and dramatically influenced by computer-based music technologies in the next years. Systems get more robust and flexible every day, a multitude of platforms is available, and there is a growing interest for pushing forward research in the field. Nonetheless, the community still faces many challenges in terms of future research directions, many of them pointing out to the imminent need for collaboration between different fields and communities. Music technologies swiftly evolve and consequently, the way people interact with music. In the same manner, music education and learning systems have to evolve and take advantage of the many possibilities provided by new technologies.

7 Acknowledgements

The Thuringian Ministry of Economy, Labour and Technology supported this research by granting funds of the European Regional Development Fund (ERDF) to the project Songs2See, enabling transnational cooperation between Thuringian companies and their partners from other European regions.

References

1 J. Abeßer, C. Dittmar, and G. Schuller. Automatic recognition and parametrization of frequency modulation techniques in bass guitar recordings. In *Proceedings of the 42nd Audio Engineering Society (AES) Conference: Semantic Audio*, Ilmenau, Germany, 2011.

2 J. Abeßer, H. Lukashevich, and G. Schuller. Feature-based extraction of plucking and expression styles of the electric bass guitar. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, US*, 2010.

3 F. Argenti, P. Nesi, and G. Pantaleo. Automatic transcription of polyphonic music based on the constant-q bispectral analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1610–1630, 2011.

4 E. Benetos and S. Dixon. Joint multi-pitch detection using harmonic envelope estimation for polyphonic music transcription. *Selected Topics in Signal Processing, IEEE Journal of*, 5(6):1111 –1123, oct. 2011.

5 J. C. Brown. Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

6 A.M. Burns and M.M. Wanderley. Visual methods for the retrieval of guitarist fingering. In *Proceedings of the 2006 International Conference on New Interfaces for Musical Expression (NIME06)*, pages 196–199, Paris, France, 2006.

7 E. Cano and C. Cheng. Melody Line Detection and Source Separation in Classical Saxophone Recordings. In *12th International Conference on Digital Audio Effects (DAFx-09)*, pages 1–6, Como, Italy, 2009.

8 E. Cano, C. Dittmar, and S. Grollmisch. Songs2See: Learn to Play by Playing. In *12th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, USA, 2011.

9 A. Cont. ANTESCOFO: Anticipatory Synchronization and Control of Interactive Parameters in Computer Music. In *International Computer Music Conference (ICMC)*, Belfast, Ireland, 2008.

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

11 C. Dittmar, K. Dressler, and K. Rosenbauer. A Toolbox for Automatic Transcription of Polyphonic Music. In *2nd Conference on Interaction with Sound- Audio Mostly*, Ilmenau, 2007.

12 K. Dressler. Sinosoidal extraction using an efficient implementation of a multi-resolution fft. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06), Montreal, Canada*, pages 247–252, 2006.

13 K. Dressler. An Auditory Streaming Approach For Melody Extraction from Polyphonic Music. In *12th International Society for Music Information Retrieval Conference (ISMIR)*, ISMIR, pages 19–24, Miami, USA, 2011.

14 J. L. Durrieu, G. Richard, and B. David. An Iterative Approach to Monaural Musical Mixture De-Soloing. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 105–108, 2009.

15 V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and Objective Quality Assessment of Audio Source Separation. Technical report, Institut National de Recherche en Informatique et en Automatique, Rennes, 2010.

16 S. Ewert and M. Müller. Score-informed voice separation for piano recordings. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, Miami, USA, 2011.

17 S. Ewert, M. Müller, and R. B. Dannenberg. Towards reliable partial music alignments using multiple synchronization strategies. In *Proceedings of the International Workshop on Adaptive Multimedia Retrieval (AMR)*, Madrid, Spain, 2009.

18 R. Fiebrink. Modeling flute fingering difficulty. Technical report, The Ohio State University, 2004.

19 D. Fitzgerald. Harmonic/Percussive Separation Using Median Filtering. In *13th International Conference on Digital Audio Effects (DAFx -10)*, page 10, 2010.

20 B. Fuentes, R. Badeau, and G. Richaed. Adaptive Harmonic Time-Frequency Decomposition of Audio Using Shift-Invariance PLCA. In *Proc. of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 401–404, 2011.

21 M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano. Songle: A web service for active music listening improved by user contributions. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, Miami, USA, 2011.

22 S. Grollmisch, C. Dittmar, E. Cano, and K. Dressler. Server based pitch detection for web applications. In *AES 41st International Conference: Audio for Games*, London, UK, 2011.

23 S. Grollmisch, C. Dittmar, and G. Gatzsche. Concept , Implementation and Evaluation of an improvisation based music video game. In *Proceedings of IEEE Consumer Electronics Society's Games Innovation Conference (IEEE GIC)*, 2009.

24 M. Hart, R. Bosch, and E. Tsai. Finding optimal piano fingerings. *The UMAP (Undergraduate Mathematics and Its Applications) Journal*, 21:167–177, 2000.

25 A. Hrybyk and Y. Kim. Combined audio and video for guitar chord identification. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht, Netherlands*, pages 159–164, 2010.

26 D. Huron and J. Berec. Characterizing idiomatic organization in music: A theory and case study of musical affordances. *Empirical Musicology Review*, 4, 2009.

27 A. A. Kasimi, E. Nichols, and C. Raphael. A simple algorithm for automatic generation of polyphonic piano fingerings. In *Proc. of the 8th International Conference on Music Information Retrieval (ISMIR), Vienna, Austria*, 2007.

28 C. Kerdvibulvech and H. Saito. Vision-based guitarist fingering tracking using a bayesian classifier and particle filters. *Advances in Image and Video Technology*, pages 625–638, 2007.

29 A. Klapuri. A method for visualizing the pitch content of polyphonic music signals. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR), Kobe, Japan*, 2009.

30 A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription*. Springer Science+Business Media, 2006.

31 M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis. Normalized Cuts for Predominant Melodic Source Separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):278–290, February 2008.

32 M. Laurson, V. Välimäki, and H. Penttinen. Simulating idiomatic playing styles in a classical guitar synthesizer: Rasgueado as a case study. In *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10), Graz, Austria*, 2010.

33 D. D. Lee and H. S. Seung. Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS), Denver, CO, USA*, pages 556–562. MIT Press, 2000.

34 Y. Li and D. Wang. Separation of Singing Voice From Music Accompaniment for Monaural Recordings. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1475–1487, May 2007.

35 25 Jahre Musikspiele. *M! Games*, 5:24–25, 2009.

36 M. Müller, D. Ellis, A. Klapuri, and G. Richard. Signal processing for music analysis. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1088 –1110, oct. 2011.

37 N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka, and S. Sagayama. Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram. In *EUSIPCO*, pages 1–4, Lausanne, Switzerland, 2008.

38 T. H. Özaslan and J. L. Arcos. Legato and glissando identification in classical guitar. In *Proc. of Sound and Music Computing Conference (SMC), Barcelona, Spain*, 2011.

39 M. Paleari, B. Huet, A. Schutz, and D. Slock. A multimodal approach to music transcription. In *Proc. of the 15th IEEE International Conference on Image Processing (ICIP)*, pages 93–96, 2008.

40 R. Parncutt, J. A. Sloboda, E. F. Clarke, M. Raekallio, and P. Desain. An ergonomic model of keyboard fingering for melodic fragments. *Music Perception*, 14:341–382, 1997.

41 L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, feb 1989.

42 D. P. Radicioni. *Computational Modeling of Fingering in Music Performance*. PhD thesis, Department of Psychology, University of Torino, Italy, 2005.

43 E. Radisavljevic and P. Driessen. Path difference learning for guitar fingering problem. In *Proc. of the International Computer Music Conference (ICMC)*, 2004.

44 C. Raphael. Music Plus One and Machine Learning. In *27th International Conference on Machine Learning*, Haifa, Israel, 2010.

45 M. Ryynänen, T. Virtanen, J. Paulus, and A. Klapuri. Accompaniment Separation and Karaoke Application Based on Automatic Melody Transcription. In *IEEE International Conference Multimedia Expo*, pages 1417–1420, 2008.

46 S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama. Specmurt Analysis of Polyphonic Music Signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(3):639–650, February 2008.

47 P. Smaragdis and M. Casey. Audio/visual independent components. In *Proc. of the 4th International Symposium on Independent Components Analysis and Blind Signal Separation, Nara, Japan*, 2003.

48 P. Smaragdis, B. Raj, and M. Shashanka. A Probabilistic Latent Variable Model for Acoustic Modeling. In *Proc. of the 20th Annual Conference on Neural Information Processing Systems (NIPS)*, 2006.

49 M. Sterling, X. Dong, and M. Bocko. Pitch bends and tonguing articulation in clarinet physical modeling synthesis. In *Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2009.

50 D. R. Tuohy and W. D. Potter. Guitar tablature creation with neural networks and distributed genetic search. In *Proc. of the 19th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA-AIE06, Annecy, France*, 2006.

51 E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca. First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results, 2007.

52 T. Virtanen, A. Mesaros, and M. Ryynänen. Combining Pitch-Based Inference and Non-Negative Spectrogram Factorization in Separating Vocals from Polyphonic Music. In *ISCA Tutorial Res. Workshop Statist. pecept. Audition.*, Brisbane, Australia, 2008.

53 D. Wagner. Implementierung und Evaluation einer interaktiven Fingersatz-Animation in Musiklernsoftware. Master's thesis, Ilmenau University of Technology, 2011.

54 Y. Yonebayashi, H. Kameoka, and S. Sagayama. Automatic decision of piano fingering based on hidden markov models. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

55 K. Yoshii and M. Goto. Infinite latent harmonic allocation: A nonparametric bayesian approach to multipitch analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht, Netherlands*, pages 309–314, 2010.

56 R. Zhou and M. Mattavelli. A new time-frequency representation for music signal analysis: Resonator time-frequency image. In *Proc. of the 9th International Symposium on Signal Processing and Its Applications (ISSPA)*, pages 1–4, feb. 2007.

Human Computer Music Performance*

Roger B. Dannenberg

Carnegie Mellon University
Pittsburgh, PA 15213, USA
rbd@cs.cmu.edu

Abstract

Human Computer Music Performance (HCMP) is the study of music performance by live human performers and real-time computer-based performers. One goal of HCMP is to create a highly autonomous artificial performer that can fill the role of a human, especially in a popular music setting. This will require advances in automated music listening and understanding, new representations for music, techniques for music synchronization, real-time human-computer communication, music generation, sound synthesis, and sound diffusion. Thus, HCMP is an ideal framework to motivate and integrate advanced music research. In addition, HCMP has the potential to benefit millions of practicing musicians, both amateurs and professionals alike. The vision of HCMP, the problems that must be solved, and some recent progress are presented.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Performing arts, H.5.1 Multimedia Information Systems

Keywords and phrases Interactive performance, music processing, music signals, music analysis, music synthesis, audio, score

Digital Object Identifier 10.4230/DFU.Vol3.11041.121

1 Introduction

Human Computer Music Performance (HCMP) is a new term intended to describe an emerging practice of creating computer music systems that can perform live music in association with human performers [11]. Interactive music performance itself is not new; however, even after decades of work in this area, examples of intelligent, competent, autonomous music performance by computer are rare. Moreover, some very general problems of music listening and generation need to be solved to enable a rich practice of HCMP. This contribution describes many problems of live music performance. It is hoped that researchers will be inspired to consider these problems and perhaps solve them.

Interactive music systems to date fall into several different categories. Probably, the most extensive work has been with experimental music where there are few traditions or constraints. This has freed creators from concerns of synchronization, harmonic structure, adherence to predetermined forms, etc. Instead, the focus can be on interactivity, gestural control, algorithmic composition, and new synthesis techniques, which have all advanced greatly over several decades [43, 46].

Another area of focus is score following and computer accompaniment [15]. These systems assume that music details are predetermined by the composer, and the main interactive task is synchronization. Typically, computer accompaniment systems have no "understanding" or representation of music theory, structure, or form, and there is no need to generate music other than to play predetermined notes or sounds. One might say computer accompaniment

* This work was partially supported by the National Science Foundation.

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 121–134

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

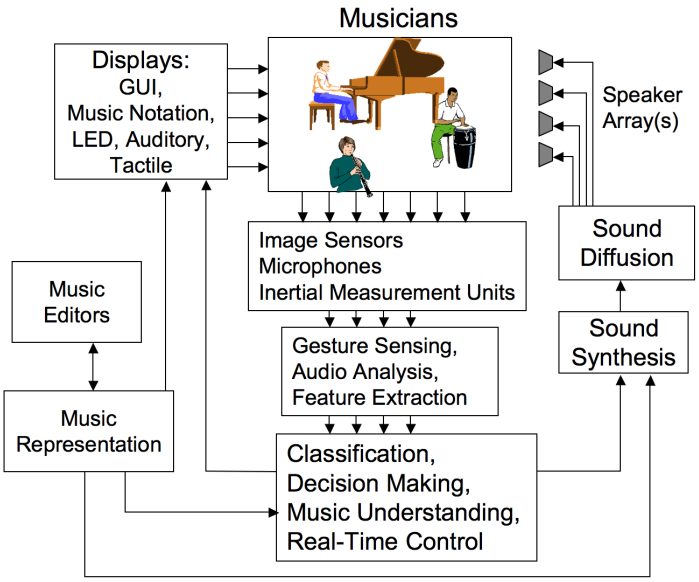


Figure 1 A high-level view of a Human Computer Music Performance (HCMP) system.

has been successful because all of the effort can be focused on two clear problems: (1) real-time alignment of a performance to a score, and (2) musically adjusting the playback of an accompaniment to synchronize to another player.

A limitation of computer accompaniment is that real-time alignment is not always useful or applicable to music synchronization. A common situation in popular music is that there is no detailed score to which a performance can be compared and aligned. For example, lead sheets may notate only chords. Even where a melody is notated in detail, it is often understood that the rhythms can be interpreted freely. Alignment to a score that is freely interpreted does not give very useful information about musical tempo and location. In these cases, a very different approach must be taken to synchronize musicians.

"Popular music" (for lack of a better term) is defined here to be music with a generally steady tempo, clear structures of sections, phrases, harmony, melody, and meter, and usually a substantial amount of improvisation. Synchronization in popular music tends to be based on beats and measures rather than on a score. The score is likely to be imprecise (as in a lead sheet that only specifies chords and melody), and even when the score suggests every note to play, musicians are usually free to interpret the music, adding their own chord voicings, strumming patterns, syncopations, etc. Because of the difficulty of synchronizing machines to humans, it is common for humans to synchronize to machines or to the playback of fixed media as in Karaoke performance, street musicians playing along with backing tracks, and practicing with Music Minus [36] recordings.

At present, HCMP is more of a vision than a practice. The vision is for autonomous computer systems to play the role of a skilled musician in a live performance of popular music.[1] Figure 1 gives a high-level view of a complete HCMP system. Musicians are sensed not only by machine listening but also through a variety of sensors and interfaces. Musical

[1] The term HCMP is really applicable to all forms of live music performance involving humans and computers. If the term is used more broadly, then we may need more specific terms, e.g. HCMP_{PM} for popular music, HCMP_{SF} for synchronization by score following, etc. I will simply use HCMP for now.

decision making relates real-time music performance to music representations that could be very specific or only musical sketches. Music is generated, synthesized, and diffused via loudspeakers into the performance space. Non-audio feedback in the form of displays, tactile feedback, and music notation are also generated to communicate with the live performers. Editors are also available to create and alter music for the HCMP system. Realizing the vision of HCMP will require solutions to a number of problems:

- Synchronization in beat-based music is largely unexplored from the standpoint of implementing systems with musical competence. Of course, automatic beat tracking has been studied, but highly reliable and robust systems do not exist. Studies of human synchronization in music are found in the literature of performance analysis, for example [41].
- Communication among musicians is especially important in popular music where the score may only be a sketch and where performers are generally free to alter the form of the music in mid-performance.
- Musicians plan performances based on very abstract representations of the music. For example, "I'll solo and you come in on the bridge" is almost a complete recipe for performing a ballad, but this is possible only because the musicians have shared conventions for describing music structure and organizing performances.
- Musicians transform some representation of music into live sound. This may involve the composition of parts, e.g. writing a bass line given a chord progression. Once notes and phrases are determined, they must be synthesized musically or perhaps performed acoustically by a robot. Finally, sounds must be diffused into the performance space, ideally in a way that conveys the impression of live performance rather than the mere playback of a recording.

After a short discussion of related work, these problems are considered one-by-one in the following sections. We conclude with a discussion of the possible impact HCMP research can have on the future of music.

2 Related Work

As mentioned above, most of the work to date on interactive computer music addresses problems of experimental contemporary art music and Western art music, but ignores the problems of popular music or music with a steady tempo, where scores are not so helpful and where timing must be very precise. In the commercial world, Ableton Live [1] provides a powerful interface for beat-based music production and control, and it provides some real-time time-stretching and tempo adjustment capabilities, but it is not meant to function as a virtual musician. Robertson and Plumbley's B-Keeper system [42] extends Ableton Live with a real-time beat tracker and user interface so that a user can synchronize music to a live drummer. This system implements some components of an HCMP system, but does not address other issues.

Conducting systems [5, 8, 13, 29] are closely related to HCMP, and differ mainly in that they assume a dedicated person (the conductor) to give commands to the computer. Conducting is certainly an interesting way to synchronize computer performers with humans in a live performance. Conducting systems have been used in public performances, including some controversial performances of traditional works where electronics was used in place of an acoustic orchestra [19].

In some conducting systems, all of the music is assumed to be generated by the computer, so cues and synchronization are not critical issues. Many conducting systems have focused on issues of adjusting tempo in real time and performing real-time time stretching of audio and video in live interactive performance of classical music.

On the other hand, popular music performance with human musicians raises the issues of accurate synchronization, and how to handle cuts, repeats, and other changes in form. Another issue is that conducting systems require a human conductor. Even if there is already a conductor, it is common to add another conductor specifically to operate the computer system. This is important because any technical "solution" that requires the full-time efforts of a person has to be compared to the possibility of simply adding another human musician to the existing ensemble. In contrast to a conducting system, an autonomous virtual musician seems to be a better solution for a small group playing popular music.

3 Beat Tracking, Tempo, and Synchronization

The irony of working with "steady beat" music is that in live performance, the tempo is never truly steady. Variations of 5 to 10 percent over time periods on the order of one minute are to be expected, and fairly sudden tempo changes are common as well. In spite of this variation, "steady beat" music is, for the most part, very steady, and the predictability of beat times is crucial to synchronization, given that all the parts may be improvised and therefore unpredictable. HCMP systems need to identify beats accurately and reliably. This might be achieved with a combination of automatic beat tracking software [7, 16, 17, 23, 24, 32, 35] and gestural sensors such as foot pedals or accelerometers.

Automation is desirable but tends to be not so reliable in two ways. First, automatic beat trackers often make serious mistakes, losing track of the beat altogether. Although the literature often implies that beat tracking is largely solved, even state-of-the-art beat trackers fail too often to be used in live performance. The second problem is that automatic beat tracking precision is not high. One might expect that when beat trackers work, they are synchronized to audio features such as snare drum hits, which should be easy to detect, reliable, and precisely timed. In reality, music audio often contains events that are slightly offset from the true beat times and which result in inaccuracies. It seems that humans and automatic beat trackers are using very different processes to identify beats, and these processes are not always consistent.

Human input through tapping is more robust that beat-tracking, but tapping can be a distraction to musicians. Also, musicians who are distracted by performance tasks can tap with inadvertently large skews between the tap times and the true beat times. As with automatic beat tracking, there are precision problems with tapping, especially foot tapping, which otherwise is one of the most reliable and least obtrusive ways of getting beat information from live performers.

In addition to identifying beats, HCMP systems need to know how beats relate to the overall music structure. An important level of structure is the measure or bar. These groups of 4 beats (typically) are the points of transitions such as chord changes and phrase beginnings. By determining measure boundaries, HCMP systems can better interpret cues or signals from humans. These cues are often ambiguous at the beat level but usually refer to the nearest measure boundary. Robustly detecting measure boundaries is an interesting real-time music analysis problem [27, 38].

4 Human Computer Communication

While computer accompaniment systems and beat trackers focus on extracting information from music audio, much of the interaction in a music performance is external to the audio channel. Musicians give visual signals, make eye contact, and use body gestures to commu-

nicate music information, including cues for synchronization. Not all signals are explicit. For example, as a trumpet player, I can usually tell when a fellow player is about to play by noticing when the trumpet is lifted and when the player takes a breath. I may not even be consciously aware of this communication, but it exerts a strong influence that may help me to take corrective action if I am not synchronized to the other players.

Similarly, HCMP systems should develop simple and natural communication channels between computer and human musicians. In fact HCMP is intended to sound a bit like HCI–Human Computer Interaction–and HCMP research needs to build upon HCI techniques and approaches [40]. This is a rich area for research, especially given the many possibilities of sensors and computer processing. We need to explore real-time interaction techniques for performing musicians as well as real-time displays that allow computers to communicate with and give cues to humans.

Nicolas Gold and I [22] have identified multiple classes of cues in an effort to describe musical communications more systematically. A *Static Score Position Cue* communicates the current position in the score to correct synchronization problems. An *Intention Cue* is used to indicate the direction of the performance when there are options. For example, "this is the last time we will repeat this section." A *Voicing/Arrangement Cue* is used to modify the performance, for example by telling a player to begin playing, to play louder, or to play more notes.

We have developed a music notation-based interface for HCMP in my lab [30]. (See Figure 2.) The music notation system, inspired by earlier work [10], is bi-directional: The computer can display its location by highlighting bar lines in the music to confirm to the human that it is in the right location. On the other hand, the human can touch or click on locations in the score to give cues or to tell the computer where to begin in a rehearsal.

Other modes of interaction are also possible. For example, we have developed a small touch sensor that can be worn on the finger and used to give cues while playing a musical instrument. Nicolas Gold wrote software to interpret several different free hand cues using a Kinect [33]. I have used foot tapping at half tempo (cut time) to indicate tempo and 4 taps at full tempo to cue the beginning of some music. The possibilities seem endless.

5 Music Structure

Music structure and representation has received much attention, but HCMP seems to raise some unique questions. The main idea is that music often includes repetition and hierarchical structure. For example, a popular song form can be described as "AABA," indicating that the first section (usually 8 measures) is repeated, followed by a contrasting "B" part, and then the "A" part is repeated again at the end. In practice, sections are usually varied slightly from one instance to the next. An interesting challenge for music processing is to detect music structure automatically from audio [2, 14, 39] or symbolic scores [28].

Popular music scores exist in multiple forms, and coordinating these somewhat informal objects systematically will require careful design of models, representations, and interfaces [22]. The typical musical score uses repeat signs and other constructs to make the notation more compact. Often, there are exceptions where the music is played differently on each repetition. Sometimes, the exceptions are handled in a standard way, such as first and second ending notation, but often there are informal annotations, e.g. "Play 1x only." We call this the *static* score, in analogy to static computer programs.

When a static score is "executed" or unfolded to create a linear sequence of events, the fact that repetitions are different and occur at different times can be represented directly,

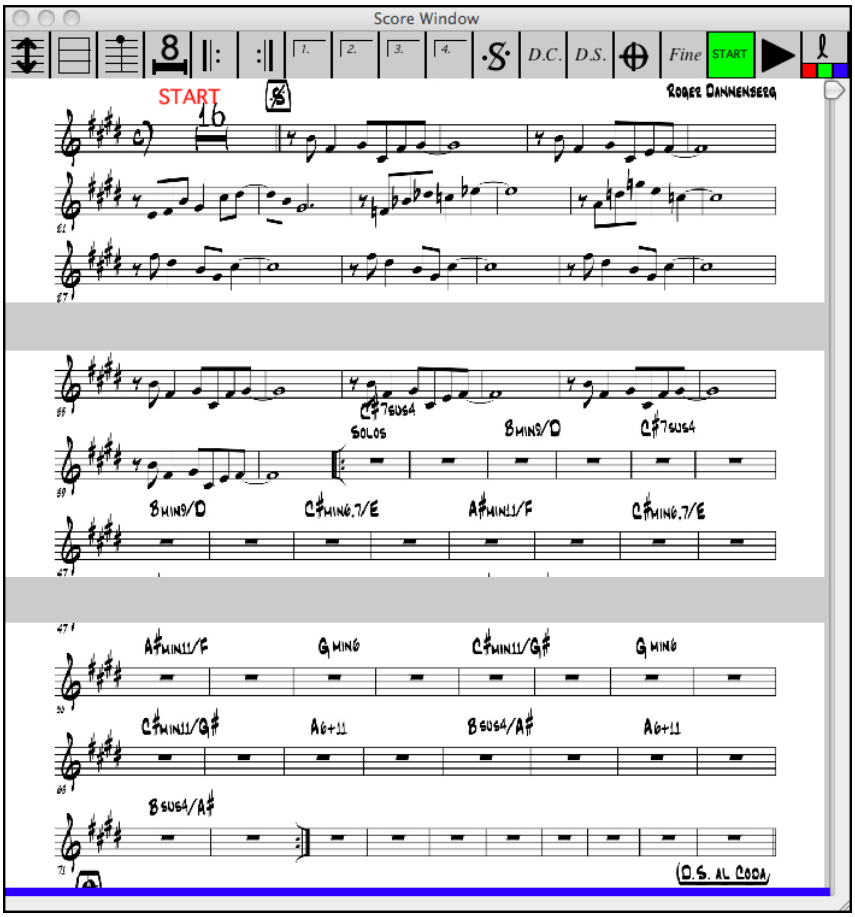


Figure 2 A bi-directional interactive music interface for HCMP. The computer can scroll music and highlight its current position. The human can point to locations to give cues, correct synchronization problems, or indicate starting locations in rehearsals.

albeit with greater redundancy. We call this the *dynamic score* in analogy to the dynamic (run-time) execution sequence of a program. The mapping between static scores and dynamic scores is complicated by non-determinism. For example, a repeat may be marked "ad lib," meaning that the number of repetitions is to be determined at performance time. Thus, the dynamic score cannot be fully determined until performance time.

However, dynamic scores are not just ephemeral traces of a performance. An audio recording corresponds to the dynamic score as does a MIDI sequence. If a performance consists of humans reading static scores and an HCMP system playing from a MIDI file and an audio file, then clearly the static and dynamic representations must be reconciled. See [21] for related work.

In popular music, scores are often informal, and musicians often create informal plans that do not match the implied plan of the score. For example, musicians might decide to play intro, verse, chorus, chorus, ending, even if the score shows a second verse. In popular music, it is accepted practice to alter the structure in this way. Musicians might even decide to change the key of the second chorus. We call these informal plans "arrangements." An HCMP system must be able to access information from a static score, a dynamic score, and an arrangement in order to make a performance plan that is consistent with the intentions

of human musicians. It must be easy for human musicians to make and communicate arrangements in a few seconds or even during the performance, because these decisions are often made on the spot during a performance. Building simple tools to manage music where these fairly abstract concepts come into play is also a challenge.

6 Music Generation, Synthesis, and Diffusion

There are many ways that computer musicians can actually produce sound. The simplest technique is to use MIDI scores and conventional synthesizers. This approach may work very well in certain situations, but it assumes there is a detailed score including dynamics and expressive timing, and that the instrument can be synthesized adequately. An alternative is to play back prerecorded audio using time-stretching techniques to adjust the tempo of the recording and synchronize it to the live performers [12]. This works well, but it requires a lot of time and effort to prepare and it assumes that live musicians are available to perform the necessary parts in the recording session. Current synthesis methods for many instruments are not very convincing, and even good synthesis methods often require such careful control that satisfactory results are very hard to obtain. Progress is being made in research systems, e.g. [18, 37], and commercial systems, e.g. [31, 34, 44], but HCMP could benefit from new research in synthesis methods. Ideally, one would like to render a score into a convincing performance including musical phrasing, stylistically appropriate timing and articulation, dynamics, and vibrato.

A common task for a popular music performer is to play according to a "lead sheet" or "chord chart," which specifies the key, harmony and structure of the music (and sometimes the melody), but few if any other details. From this information, a drummer can create an appropriate rhythm that matches the structure of the song, a bass player can provide a rhythmic and harmonic foundation, a keyboard or guitar player can play chords according to the harmony, and other musicians can harmonize the melody or improvise a solo. Writing a new song might be considered a highly creative and difficult task, but creating a bass part or playing piano to accompany a singer is a routine task for a working musician. In fact, most musicians can create very musical parts without errors in real time as they read a lead sheet. In the field of popular music performance, many musicians are actually more comfortable improvising parts from a lead sheet than reading conventional music where every note is indicated explicitly. Thus, creating musical parts from a sketch such as a lead sheet is a fundamental skill that should be implemented by an HCMP system. This is especially important when human performers do not already play the instrument to be played by HCMP and therefore lack the skill or experience to compose the part.

Programs such as Band-in-a-Box [26] perform the music-from-lead-sheet task already, but do not give the user too much control over music generation. Instead, Band-in-a-Box offers a wide variety of styles from which the user can select. It seems that research into machine learning, musical analogy, music similarity, and models of musical style can lead to more flexible and controllable music generation.

Finally, one of the problems with computer-generated sound is diffusion into an acoustic space. The one-dimensional audio signal must be converted to three-dimensional sound waves through loudspeakers, which impart their own audible radiation characteristics onto the sound. As computer and digital audio equipment prices have fallen, there has been much interest in using many audio channels to drive arrays of speakers to improve and control the diffusion of sound in two or three dimensions. Examples include linear speaker arrays to produce controlled wavefronts [4, 6], spherical speaker arrays to simulate sound sources with

frequency-dependent radiation patterns [3, 9], and our own work on convolution-based stereo panning and placement [25]. The challenge for HCMP is to completely integrate computer performance with live acoustic instruments. The techniques will depend upon the situation, but there is a need for improvements in sound reinforcement with implications for all audio interfaces.

7 Example

One substantial performance [12] using HCMP has been presented at Carnegie Mellon University. I worked with the student jazz band, with assistance from David Pellow, the director, and John Wilson, an arranger. We set out to create a string orchestra that could play along with a live jazz band. Our musical goal was that the strings plus jazz band sound as good as possible. We decided to emphasize practical considerations and reliability over cutting-edge research, and our experience has helped to formulate some of the problems and approaches suggested above.

John Wilson was commissioned to write for jazz band and strings. We decided to structure the piece so that the string players do not play continuously, but instead play during a number of segments of the music. The human rhythm section (drums, bass, and piano) plays continuously throughout the piece. This allowed for interplay between the band and the strings. It also had functional purposes: Because the strings were silent at many times, each entrance could be cued separately. If anything went wrong, there would soon be an opportunity to make another entrance. In addition, the sectional nature allowed for efficient recording and editing.

To the computer, the string parts are just sounds that need to be cued to begin on a particular beat and synchronized to the following beats. When each sound ends, the system prepares to play the next. Synchronization is handled in the simplest way imaginable. A foot pedal is used to tap beats (in cut time, about 85 taps per minute) to establish the tempo and throughout the performance. A small keyboard is used to cue entrances.

One thing we learned is that some new listening skills are required. In one rehearsal, the tapper (a skilled percussionist) started listening to the string section under control and naturally started to tap along with the strings rather than the band. Once the band was ignored, the tapper and strings started to drift away from the live band. As soon as the strings were obviously out of synchronization, the tapper resynchronized to the live band, but I had written code to reject "spurious" taps that fell more than 30% of a beat from the expected beat time, and the corrective taps were ignored. (Since then, we have disabled the "spurious" tap rejection feature.)

Sound generation is based on the pitch-synchronous, overlap-add (PSOLA) [45] approach to time stretching. The requirement for real-time performance, continuous (every tap) update, latency compensation, and synchronization across multiple (20) channels led to some innovative implementation details. Most PSOLA systems are designed to time stretch by a given factor over a given time span. PSOLA works by inserting or deleting whole pitch periods, thus the operation of PSOLA is not really continuous stretching but an approximation that is quantized to pitch periods. We have 20 separate channels, each with a single instrument and its own set of pitch periods. Due to the quantization, it is difficult to predict the exact duration of input audio that will be consumed, and there is the possibility that stretched tracks will lose synchronization with the accumulation of quantization errors. To ensure that all tracks remain synchronized, we use a feedback mechanism: The overall control system adjusts the global stretch factor so that the mean audio file position will

Figure 3 An HCMP performance. The live jazz ensemble is complemented by a 20-piece virtual string orchestra played over an array of loudspeakers visible behind the band.

synchronize with the live band. Then, the per-track stretch factor is adjusted slightly for each track to drive the track's audio file position toward the global mean.

Finally, sound diffusion is based on multiple (8) speaker systems arranged across the stage (see Figure 3). Each of the 20 input channels represents one close-miked string (violin, viola, or cello). Each instrument channel is directed to only one speaker. Rather than a homogenized orchestra sound spread across many speakers, we have individual instrument sounds radiating from multiple locations and mixing in the room as with an acoustic ensemble. The results were so convincing that an audio engineer wanted to know what reverberation plug-in we used for the recordings, yet the recordings were actually dry and all sense of "stereo" and reverberation resulted from the diffusion scheme.

8 Conclusions

From a scientific standpoint, Human Computer Music Performance (HCMP) offers a framework to organize, motivate, and coordinate an array of interesting research efforts. It suggests that we study music and music performance from many points of view, developing new techniques for sensing music beats, tempo, and structure, as well as new ways for musicians to communicate music intentions, especially to computer-based performers. The fundamental problems underlying HCMP are general problems of Music Understanding, thus there are broad implications. As the "Multimodal Music Processing" theme of this book suggests, this work is truly multimodal, dealing with various levels of discrete and symbolic scores, music performance data such as MIDI, music audio, graphical displays, gesture sensors, and other forms of musical communication. The models and analysis techniques introduced will have applications in other music-related studies such as music information retrieval, music theory, and music cognition.

Digital sound synthesis has been an object of study for half a century, and has connections to auditory perception, acoustics, digital signal processing, speech synthesis, and mathematics. HCMP challenges us to investigate new techniques for time-stretching and pitch shifting as well as to consider the importance of sound diffusion in the perception of music synthesis quality.

More broadly, HCMP addresses complex, real-time cooperative tasks. New interfaces are

needed to coordinate computers and humans with a minimal amount of explicit or manual control. This could have implications for other human-computer interaction scenarios such as driving and piloting, directing disaster relief, or complex mission control where tasks must be delegated and coordinated. HCMP problems suggest an integrated approach that combines machine learning with human factors studies to create reliable interfaces; advances in this area should have many applications beyond music.

Music making is practiced in the majority of households in the United States. The National Association of Music Merchants reported that industry retail sales in 2006 were about 8 billion dollars in the U.S. This includes sales of over 5 million musical instruments, but does not include music education, music recordings, or music performance. Thus, the potential societal impact of effective new music technology is enormous. HCMP seems to be an application area where recent advances in music understanding and music information processing can be leveraged to benefit millions of people. Producing results that are extremely practical and useful is not just an altruistic project for researchers. By integrating academic research to create a practice of popular music-making, the research community stands to gain greater recognition and support from society.

Another motivation for research in Human Computer Music Performance is purely artistic. One could criticize HCMP as an effort to further reduce and automate popular music, which is already formulaic. Would it not be better to devote efforts to experimental music and new art forms? My hope is to leverage the conventions, opportunities, and sheer numbers in popular music to obtain a widespread practice of HCMP. I am convinced that if this succeeds, at least a few artists in a million will invent some truly creative uses for HCMP technology that transcend existing musical practice. Thus, HCMP could be an important path by which technology shapes the future of music.

9 Acknowledgements

Thanks to Ryan Calorus, who implemented our first experimental music display, Nicolas Gold for valuable discussions, and my students Dalong Cheng, Zeyu Jin, Dawen Liang, Jiuqiang Tang, and Gus Xia. Our first performance system and the music display work were supported by Microsoft Research and the Carnegie Mellon School of Music. Zplane kindly contributed their high-quality audio time-stretching library [20] for our use, and David Pellow, Riccardo Shulz, and John Wilson made essential musical contributions to our performance with the Carnegie Mellon Jazz Ensemble. Current work is supported by the National Science Foundation under Grant No. 0855958.

References

1 Ableton AG. *Ableton Reference Manual*, 2010.

2 Jean-Julien Aucouturier and Mark Sandler. Finding repeating patterns in acoustic musical signals: Applications for audio thumbnailing. In *AES22 International Conference on Virtual, Synthetic and Entertainment Audio*, pages 412–421. Audio Engineering Society, 2002.

3 Rimas Avizienis, Adrian Freed, Peter Kassakian, and David Wessel. A compact 120 independent element spherical loudspeaker array with programmable radiation patterns. In *Proceedings of the AES 120th Convention*, 2006. Paper No. 6783.

4 Marije A. J. Baalman. Application of wave field synthesis in electronic music and sound installations. In *The ICMC 2004 Proceedings*, pages 692–698, San Francisco, 2004. The International Computer Music Association.

5 Takashi Baba, Mitsuyo Hashida, and Haruhiro Katayose. "VirtualPhilharmony": A conducting system with heuristics of conducting an orchestra. In *Proceedings of the 2010 Conference on New Interfaces for Musical Expression (NIME 2010)*, pages 263–270. ACM Press, 2010.

6 A. J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America*, 93(5):2764–2778, 1993.

7 Paul Brossier. *Automatic Annotation of Musical Audio for Interactive Applications*. PhD thesis, Department of Electronic Engineering, Queen Mary, University of London, 2006.

8 Bernd Bruegge, Christoph Teschner, Peter Lachenmaier, Eva Fenzi, Dominik Schmidt, and Simon Bierbaum. Pinocchio: conducting a virtual symphony orchestra. In *ACE '07 Proceedings of the international conference on Advances in computer entertainment technology*, pages 294–295. ACM, 2007.

9 Perry R. Cook, Georg Essl, Georgos Tzanetakis, and Dan Trueman. N >> 2: Multi-speaker display systems for virtual reality and spatial audio projection. In *Proceedings of the International Conference on Auditory Display (ICAD)*, 1998.

10 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: Special Issue on Music Digital Libraries*, 2011, to appear.

11 Roger B. Dannenberg. New interfaces for popular music performance. In *Seventh International Conference on New Interfaces for Musical Expression: NIME 2007*, pages 130–135, New York, NY, June 2007. New York Univ., ACM Press.

12 Roger B. Dannenberg. A virtual orchestra for human-computer music performance. In *Proceedings of the 2011 International Computer Music Conference*, pages 185–188. The International Computer Music Association, 2011.

13 Roger B. Dannenberg and Ken Bookstein. Practical aspects of a midi conducting program. In *Proceedings of the 1991 International Computer Music Conference*, pages 537–540. International Computer Music Association, October 1991.

14 Roger B. Dannenberg and Masataka Goto. *Music Structure Analysis from Acoustic Signals*, volume 1, pages 305–331. Springer Verlag, 2009.

15 Roger B. Dannenberg and Christopher Raphael. Music score alignment and computer accompaniment. *Communications of the ACM*, 49(8):38–43, August 2006.

16 Matthew E. P. Davies and Mark D. Plumbley. A spectral difference approach to downbeat extraction in musical audio. In *Proceedings of the 14th European Signal Processing Conference (EUSIPCO 2006)*, 2006.

17 Daniel P. W. Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007.

18 Gianpaolo Evangelista and Fredrik Eckerholm. Player-instrument interaction models for digital waveguide synthesis of guitar: Touch and collisions. *IEEE Transactions on Audio, Speech and Language Processing*, 18(4):822–832, 2010.

19 Shirley Fleming. The virtual orchestra. *American Record Guide*, 66(6), Nov./Dec. 2003.

20 Tim Flohrer. *Elastique 2.0 SDK Documentation*. zplane.development, 2007.

21 Christian Fremerey, Meinard Müller, and Michael Clausen. Handling repeats and jumps in score-performance synchronization. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, pages 243–248, Utrecht, Netherlands, 2010.

22 Nicolas Gold and Roger B. Dannenberg. A reference architecture and score representation for popular music human-computer music performance systems. In *Proceedings of the 2011 International Conference on New Interfaces for Musical Expression (NIME11)*, 2011.

23 Masataka Goto. An audio-based real-time beat tracking system for music with or without drum sounds. *Journal of New Music Research*, 30(2):159–171, 2001.

24 Peter Grosche and Meinard Müller. Extracting predominant local pulse information from music recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1688–1701, 2011.

25 William D. Haines, Jesse R. Vernon, Roger B. Dannenberg, and Peter Driessen. Placement of sound sources in the stereo field using measured room impulse responses. In *Proceedings of the 2007 International Computer Music Conference, Volume I*, pages I–496–499, San Francisco, August 2007. The International Computer Music Association.

26 PG Music Inc. Band-in-a-box 2012. `http://www.pgmusic.com` (retrieved 07.03.2012), 2012.

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

28 Olivier Lartillot. A musical pattern discovery system founded on a modeling of listening strategies. *Computer Music Journal*, 28(3):53–67, 2004.

29 Eric Lee, Thorsten Karrer, and Jan Borchers. Toward a framework for interactive systems to conduct digital audio and video streams. *Computer Music Journal*, 1(30):21–36, Spring 2006.

30 Dawen Liang, Guangyu Xia, and Roger B. Dannenberg. A framework for coordination and synchronization of media. In *Proceedings of the 2011 International Conference on New Interfaces for Musical Expression (NIME11)*, pages 167–172, 2011.

31 Eric Lindemann. Synful. `http://www.synful.com` (retrieved 07.03.2012), 2012.

32 M. F. McKinney, D. Moelants, M. E. P. Davies, and A. Klapuri. Evaluation of audio beat tracking and music tempo extraction algorithms. *Journal of New Music Research*, 36(1):1–16, 2007.

33 Microsoft. Xbox 360 + kinect. `http://www.xbox.com/kinect` (retrieved 06.03.2012), 2012.

34 Modartt. Pianoteq - product catalog. `http://www.pianoteq.com/catalog` (retrieved 07.03.2012), 2012.

35 João Lobato Oliveira, Fabien Gouyon, Luis Gustavo Martins, and Luis Paulo Reis. Ibt: A real-time tempo and beat tracking system. In *Proceedings of the 11th International Conference on Music Information Retrieval*, pages 291–296, 2010.

36 Music Minus One. Music minus one home page. `http://www.musicminusone.com` (retrieved 06.03.2012), 2012.

37 Jyri Pakarinen. Physical modeling of flageolet tones in string instruments. In *Proceedings of the 13th European Signal Processing Conference*, pages 4–8, 2005.

38 Hélène Papadopoulos and Geoffroy Peeters. Joint estimation of chords and downbeats from an audio signal. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):138–152, 2011.

39 Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-based music structure analysis. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, pages 625–636, Utrecht, Netherlands, 2010.

40 Jenny Preece, Yvonne Rogers, Helen Sharp, David Benyon, Simon Holland, and Tom Carey. *Human-Computer Interaction: Concepts and Design.* Addison-Wesley Pub. Co., 1994.

41 Bruno H. Repp, Justin London, and Peter E. Keller. Production and synchronization of uneven rhythms at fast tempi. *Music Perception*, 23:61–78, 2005.

42 Andrew Robertson and Mark Plumbley. B-keeper: A beat-tracker for live performance. In *New Interfaces for Musical Expression*, pages 234–237, 2007.

43 Robert Rowe. *Interactive Music Systems.* MIT Press, Cambridge, MA, 1993.

44 Samplemodeling. Sample modeling products. `http://www.samplemodeling.com/en/products.php` (retrieved 07.03.2012), 2012.

45 Norbert Schnell, Geoffroy Peeters, Serge Lemouton, Philippe Manoury, and Xavier Rodet. Synthesizing a choir in real-time using pitch synchronous overlap add (PSOLA). In *Proceedings of the 2000 International Computer Music Conference*, pages 102–108, 2000.

46 Tod Winkler. *Composing Interactive Music: Techniques and Ideas Using Max*. MIT Press, Cambridge, MA, 2001.

User-Aware Music Retrieval

Markus Schedl[1], Sebastian Stober[2], Emilia Gómez[3], Nicola Orio[4], and Cynthia C. S. Liem[5]

1 **Department of Computational Perception**
Johannes Kepler University, Linz, Austria
markus.schedl@jku.at
2 **Data & Knowledge Engineering Group**
Otto-von-Guericke-Universität, Magdeburg, Germany
stober@ovgu.de
3 **Music Technology Group**
Universitat Pompeu Fabra, Barcelona, Spain
emilia.gomez@upf.edu
4 **Department of Information Engineering**
University of Padova, Italy
orio@dei.unipd.it
5 **Multimedia Information Retrieval Lab**
Delft University of Technology, the Netherlands
c.c.s.liem@tudelft.nl

Abstract

Personalized and user-aware systems for retrieving multimedia items are becoming increasingly important as the amount of available multimedia data has been spiraling. A personalized system is one that incorporates information about the user into its data processing part (e.g., a particular user taste for a movie genre). A context-aware system, in contrast, takes into account dynamic aspects of the user context when processing the data (e.g., location and time where/when a user issues a query). Today's user-adaptive systems often incorporate both aspects.

Particularly focusing on the music domain, this article gives an overview of different aspects we deem important to build personalized music retrieval systems. In this vein, we first give an overview of factors that influence the human perception of music. We then propose and discuss various requirements for a personalized, user-aware music retrieval system. Eventually, the state-of-the-art in building such systems is reviewed, taking in particular aspects of *similarity* and *serendipity* into account.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems, I.5 Pattern Recognition

Keywords and phrases user-aware music retrieval, personalization, recommendation, user context, adaptive systems, similarity measurement, serendipity

Digital Object Identifier 10.4230/DFU.Vol3.11041.135

1 Introduction

Multimodal music processing and retrieval can be regarded as subfields of music information research (MIR), a discipline that has substantially gained importance during the last decade. Multimodality can be recognized at several levels in MIR, for example, different modalities to access music collections (query-by-example, direct querying, browsing, metadata-based search, visual user interfaces) or different representations of music items themselves – score sheet, symbolic MIDI, digital audio waveform, or textual lyrics, just to name a few.

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 135–156

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

In this article, multimodality relates to the integration of *various knowledge sources* in music processing systems. A key source of knowledge is given by aspects linked to the user and his or her usage of the system, which is the focus of the present study. The article at hand hence gives an overview of the state-of-the-art in modeling and determining properties of music and listeners using features of different kinds. These features all relate to how music is perceived by humans. First, a broad categorization of such features is presented in Section 2. Also references to related work on extracting and processing the respective features is given for each feature category. Subsequently, various research endeavors and directions deemed to be important by the authors for the future of personalized, multimodal music retrieval are presented. More precisely, we present a set of requirements important for user-aware music retrieval systems in Section 3. Two vital prerequisites to build user-aware music retrieval applications, such as personalized music recommender systems or user-adaptive browsing interfaces, are first *elaborating similarity measures* that are capable of revealing similarity relations as perceived by humans and second provide a *serendipitous experience* to the user. In order to develop the mentioned, sophisticated similarity measures, we need methods that capture musical similarity at different levels using different modalities, for example, timbre, rhythm, harmony, lyrics, or co-listening information. A review of the state-of-the-art in building such adaptive similarity measures is presented in Section 4. The latter requirement, ensuring a certain degree of serendipity in retrieval results, necessitates to take into account various user-dependent factors. For example, it is important for a serendipitous system to have information about the user's music taste and preference, where taste refers to a long-term inclination and preference describes a rather short-term, situation-dependent affection. Both are likely to change over time, although taste usually changes only gradually and at a slower rate than preference. More details on serendipity aspects in personalized music retrieval are given in Section 5. Finally, in Section 6, we draw conclusions and indicate some directions for future research.

2 Computational Aspects of Music Perception and Similarity

Developing computational features that encode knowledge on how we humans perceive music is one of the grand challenges in MIR. It is a particular endeavor for various reasons. Among others, music perception is very subjective and influenced for example by the listener's music preferences, but also highly dependent on his or her musical training as well as social and sociographic background. Moreover, perceptually relevant features may be extracted from very different media and representations of music, which describe a wide variety of aspects. Media encoding music or music-related data range from score sheets to digital audio files and from textual lyrics to images of cover artwork. Which of these multimodal aspects influence human perception of music, in which way and to which extent is still an open research question.

Computational music features can be broadly categorized into three classes, according to the authors: *music content*, *music context*, and *user context*, cf. Figure 1.

2.1 Music Content

In traditional MIR, features extracted by applying signal processing techniques to audio signals were dominant. Such features are commonly denoted as *signal-based*, *audio-based*, or *content-based*. In addition to audio signals, the music content may be described by various other modalities, such as handwritten or digitized score, or video clips.

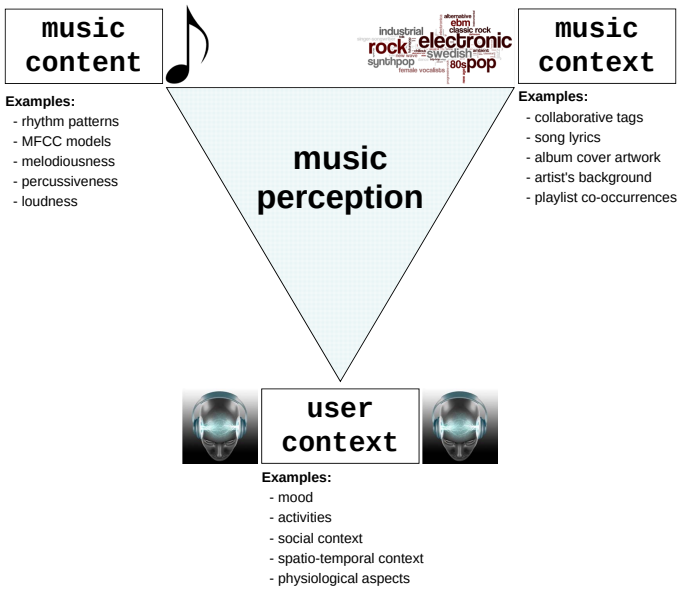


Figure 1 Feature categories to describe music.

Thorough overviews of common extraction techniques are presented in [17, 28, 65]. Music content-based features may be low-level representations that stem directly from the audio signal, for example Mel Frequency Cepstral Coefficients (MFCCs) [54], zero-crossing rate [29], amplitude envelope [15], bandwidth and band energy ratio [52], spectral centroid [82], fundamental frequency or chroma features [11]. As mentioned in [28], most low-level features do not make sense to the majority of the listeners, although they are easily exploited by computing systems.

Alternatively, content-based features may be derived or aggregated from low-level properties, and therefore represent aspects on a higher level of music understanding. Such features are often named mid-level features. Machine learning, statistical modeling and models of the human auditory system make mid-level descriptors possible, usually by gathering large sets of observations. Mid-level features usually aim at capturing either *timbral aspects* of music, which were traditionally modeled via *MFCCs* [2], *rhythmic aspects*, for example described via beat histograms [92] or fluctuation patterns [78, 69], and *tonal aspects* such as predominant melody[73], key or chord progression [27], often derived from chroma features.

Recent work aims at inferring more specific high-level concepts, meaningful to users, such as melodiousness, complexity, danceability, aggressiveness [70, 68, 90], mood [44], or genre [31]. The transition from low- or mid-level descriptors to high-level descriptors requires bridging the semantic gap. According to [28], high-level or semantic feature extractors require to include an induction procedure that has to be carried out by means of a user model, and not only a data model as in the case of mid-level descriptors.

2.2 Music Context

The *music context* can be described as all information relevant to the music item under consideration, albeit not directly extractable from the music manifestation itself. For example,

the meaning of a song's lyrics [40, 36], the political background of the musician, or the geographic origin of an artist [30, 81, 80] are likely to have a strong impact on how music is perceived and interpreted, but are not manifested in the signal.

An overview of the state-of-the-art in *music context*-based feature extraction (and similarity estimation) can be found in [76]. The majority of the approaches covering the music context are strongly related to *Web content mining* [53] as the Web provides contextual information on music artists in abundance. For example, in [34] the authors construct term profiles created from *artist-related Web pages* to derive music similarity information. *RSS feeds* are extracted and analyzed in [18]. Alternative sources to mine music context-related data include *playlists* (e.g., radio stations and mix tapes, i.e., user-generated playlists) [3, 16, 67] and *Peer-to-Peer networks* [83, 55, 24, 96]. In these cases, *co-occurrence analysis* is commonly employed to derive similarity information on the artist- or track-level. Co-occurrences of artist names on Web pages are also used to infer artist similarity information [77] and for artist-to-genre classification [79]. *Song lyrics* as a source of music context-related information are analyzed, for example, in [56] to derive similarity information, in [45] for mood classification, and in [60] for genre classification. Another source for the music context is *collaborative tags*, mined for example from *last.fm* [43] in [25, 51] or gathered via *tagging games* [59, 91, 46].

2.3 User Context

Scientific work on MIR that takes into account aspects of the user context is still relatively sparse and covers diverse topics. It can be broadly divided into user music-seeking behavior studies, user preferences elicitation, multifaceted user and similarity models, and personalized, user-aware recommender systems.

User Music-Seeking Behavior Studies

Several MIR researchers, largely with backgrounds in library and information sciences, have devoted studies to music-seeking behavior and information requirements of users. While these studies typically are conducted on a much smaller-scaled population than usually found in engineering settings, they are detailed and give qualitative insight into real-life and every-day music behavior. Many of them strikingly point out how the reception of music is not just guided by the characteristics of the music audio signal, but is strongly influenced by multimodal influences that do not necessarily have to do with the music.

Cunningham et al. [22] conducted an ethnographic study of music searching and browsing techniques. Important findings regarding this chapter were that music shopping often was a collaborative activity, with a social function going beyond music listening, and a 'surprisingly visual' activity too, with shoppers identifying music genres that they liked through the appearance of album covers. The influential role of visual means in musical settings also appears in other user studies, e.g. Bainbridge et al. [6], in which a user-centered personal digital library is designed with the spatial hypermedia paradigm, and recently Barthet and Dixon [10], describing an ethnographic study of musicologists at the British Library. In the latter study, visualization of audio signals aided the musicologists with exploring and studying music recordings, but also could steer the users' attention towards specific details. A visual spectrogram display pointed out signal features (e.g. vibrato) that the user was not aware of, but also deemphasized sound aspects that could not be seen: "I completely forgot about the bassoon, it feels like it is unimportant now, but I was once struck by it".

Social context is a strong influence on music taste. Laplante [42] found that young adults

had a strong penchant for informal channels (e.g. friends), but a low trust of experts (e.g. music store staff). Furthermore, it was noted that music discoveries often were the result of passive rather than active search behavior – this points towards serendipitous finds, which will be discussed in the upcoming sections of this chapter.

The reasons why we remember, like or hate music also are strongly determined by context. In a study of reasons why people dislike songs [21], the factors of influence were lyrics, the earworm effect (getting a song stuck in your head without wanting this), quality of the singing voice, dislike of music videos, over-exposure of a song, pretentiousness of the performing artist, clashing taste cultures (disliking the social community associated with a certain style) and unfortunate personal associations. An extensive study by Lee [49] of natural language music queries also illustrates frequent associative notions: dormant searches get rekindled because similar thematic context settings are encountered (e.g. searching for information on a 'spooky tune' that has been used in cartoons to signify that someone has died, after hearing it being played on Halloween), and songs get a special affective meaning because they had been heard in special affective settings ("My grandfather, who was born in 1899, used to sing me to sleep with this song and I can't remember the words").

Findings from user studies as described in this paragraph have not widely been adopted in the design of MIR systems yet, but still will be very relevant when studying user context.

User Preferences Elicitation

An obvious way to obtain information about the taste, preferences and behavior of a user is context logging. However, this can pose privacy issues. In a study on users' acceptance of context logging in the context of music applications by Nürnberger and Stober [89], the authors found significant differences in the participants' willingness to reveal different kinds of personal data on various scopes. Most participants indicated to eagerly share music metadata, information about ambient light and noise, mouse and keyboard logs, and their status in instant messaging applications. When it comes to used applications, facial expressions, bio signals, and GPS positions, however, a majority of users are reluctant to share their data. As for country-dependent differences, US-Americans were found to have on overall much lesser reservations to share personal data than Germans and Austrians. One has to note, however, that the results might be biased as 70% of the 305 participants were from Germany.

An alternative to context logging is to explicitly ask the users to provide means to characterize their musical preferences. One example of this methodology is presented in [32]. This study proposes a method to automatically generate, given a provided set of preferred music tracks, an iconic representation of a user's musical preferences – the Musical Avatar. Starting from the raw audio signals, they compute a set of semantic descriptors which are mapped to the visual domain by creating a humanoid cartoony character that represents the user's musical preferences. Examples of possible avatars are provided in Figure 2. This representation of a users's musical preferences is then used to provide personalized recommendations in [13].

User and Similarity Models

One of the earliest works in user modeling for MIR is [19], where Chai and Barry present some general considerations on modeling the user in a music retrieval system. They also suggest an XML-based user modeling language for this purpose.

Zhang et al. present *CompositeMap* [100, 101], a model that takes into account similarity aspects derived from music content as well as social factors. The authors propose a multimodal

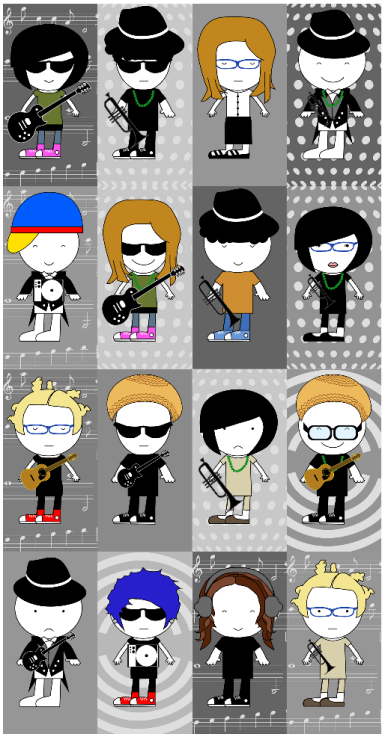


Figure 2 Examples of Musical Avatars representing the user's musical preferences [58].

music similarity measure and show its applicability to the task of music retrieval. They also allow a simple kind of personalization of this model by letting the user weight the individual music dimensions on which similarity is estimated. However, they do neither take the user context into consideration, nor do they try to learn a user's preferences.

In [63] a multimodal music similarity model on the artist-level is proposed. To this end, McFee and Lanckriet calculate a *partial order embedding* using *kernel functions*. Music context- and content-based features are combined by this means. However, this model does not incorporate any personalization strategies.

In [72] Pohle et al. present preliminary steps towards a simple personalized music retrieval system. Based on a clustering of community-based tags extracted from *last.fm*, a small number of musical concepts are derived using *Non-Negative Matrix Factorization* (NMF) [48, 98]. Each music artist or band is then described by a "concept vector". A user interface allows for adjusting the weights of the individual concepts, based on which artists that match the resulting distribution of the concepts best are recommended to the user. Zhang et al. propose in [100] a very similar kind of personalization strategy via user-adjusted weights.

Knees and Widmer present in [37] an approach that incorporates *relevance feedback* [74] into a text-based music search engine [35] to adapt the retrieval process to user preferences. The search engine proposed by Knees et al. builds a model from music content features (MFCCs) and music context features (term vector representations of artist-related Web pages). To this end, a weight is computed for each (term, music item)-pair, based on the term vectors. These weights are then smoothed, taking into account the closest neighbors according to the content-based similarity measure (Kullback-Leibler divergence on Gaussian Mixture Models of the MFCCs). To retrieve music via natural language queries, each textual query issued to the system is expanded via a *Google* search, resulting again in a term weight

vector. This query vector is subsequently compared to the smoothed weight vectors describing the music pieces, and those with smallest distance to the query vector are returned.

Nürnberger and Detyniecki present in [66] a variant of the *Self-Organizing Map* (SOM) [38] that is based on a model that adapts to *user feedback*. To this end, the user can move data items on the SOM. This information is fed back into the SOM's codebook, and the mapping is adapted accordingly.

In [99] Xue et al. present a *collaborative personalized search model* that alleviates the problems of *data sparseness* and *cold-start for new users* by combining information on different levels (individuals, interest groups, and global). Although not explicitly targeted at music retrieval, the idea of integrating data about the user, his peer group, and global data to build a social retrieval model might be worth considering for MIR purposes.

User-Aware Music Recommendation

Baltrunas et al. present a user-aware music recommender system for usage in cars [7]. They aim at learning relations between user aspects and music genres. As contextual aspects, Baltrunas et al. look into driving style, road type, landscape, sleepiness, traffic conditions, mood, weather, and time of day. Using a Web-based tool, the authors first assess in a user study which of these contextual aspects influence the preference for music of a particular genre, either in a positive of negative way. According to the study, driving style strongly influences the choice for music from the genres Blues, Classical, and Metal, whereas sleepiness seems to foster the decision for Pop, Country, and Reggae music. Furthermore, Baltrunas et al. investigate the impact of user context on user ratings and found that in most cases the awareness of a particular contextual situation had a negative effect on the ratings. The most significant (negative) influence on user ratings had the conditions "sleepy" and "traffic jam". The authors of [7] then propose a music recommendation approach that employs an extended Matrix Factorization [39] algorithm to predict item ratings. Their model includes contextual condition and genre vectors.

Bogdanov et al. [13] present a system which automatically generates recommendations from a user's musical preferences, given her/his accounts on popular online music services. Using these services, the system retrieves a set of tracks preferred by a user, and further tries to infer a semantic description of musical preferences from raw audio information. Thereafter, the system generates music recommendations, using a semantic music similarity measure.

Even though no detailed information on their approach is publicly available, *last.fm* [43] builds user models based on its users' listening habits, which are mined via the "AudioScrobbler" interface. Based on this data, *last.fm* offers personalized music recommendations and playlist generation, however, without letting the user control (or even know) which factors are taken into account. Another commercial example employing a *collaborative filtering* (CF) [14] approach can be found in *amazon.com*'s music Web store [1]. Again, no details of the exact approach are publicly available.

2.4 Further Remarks

Having presented the three basic feature categories (music content, music context, and user context), we would like to note that there is an overlap between some of these. Indeed, particular features cannot only be assigned to one group, but combine aspects of several categories. For example, song lyrics are in principal music content. However, even state-of-the-art techniques do not allow for converting sung lyrics into textual representations from the audio signal, or even to derive some kind of higher level meaning. On the other hand, several

lyrics portals on the Web (music context sources) offer such textual representations. Another example is similarity measures based on collaborative filtering. They are music context-related in the sense that the process is collaborative, however CF is used for personalizing a music recommendation to a user or a group of users, hence it takes into account the user context.

3 Important Aspects for Personalized Music Retrieval

Traditionally, evaluating music retrieval approaches focused on the concept of musical similarity, meaning that the performance of a retrieval system is judged the better the more similar the returned pieces are to a given seed. Although this is a very intuitive manner of assessment, it does not take into account that the information need of the user might be different. Indeed, for many common and popular MIR tasks, such as automated playlist generation and music recommendation, the listener does not necessarily want to be offered a list of closest matches in terms of acoustic similarity, as usually given by today's content-based music recommenders. User studies focusing on the perceived quality of automated, content-based playlist generation [71, 50] showed that playlists with items that were acoustically very similar were often deemed too perfect or homogeneous, and thus boring. In addition, users were shown to judge playlist items differently based on the amount of (metadata) information accompanying the playlist item [9, 50].
We therefore believe that a new generation of user-aware music retrieval systems should not only focus on traditional similarity scores derived via applying audio signal processing techniques, but also take other factors, including information from different modalities, into account. More precisely, such factors include the following:

Similarity

Similarity relations in various dimensions should be taken into account. One set of dimensions might be based on music properties such as rhythm, harmony, or timbre, inferred from the audio signal; another might take into account the resemblance according to other data sources, such as collaborative tags, playlist co-occurrences, or even images of album covers. A third set of dimensions might be learned from a user's listening preferences, for example, by relating certain properties of the user context to particular categories of music. To give an example, similarity could be defined as pieces that are usually listened together while a user is jogging or while being together with friends.

Moreover, the user's preferred music material should also influence the features and their relevance for similarity computation. For instance, a retrieval system focusing in classical music would need musically meaningful descriptors and similarity measures, while in a retrieval scenario of mainstream popular music timbre can be informative enough for distinguishing different types of music.

Diversity

Although the items in the results set of a music retrieval request should be similar, they should also reveal a certain degree of diversity. For example, there is the well-known "album" effect [95], i.e., due to same recording settings, tracks on one and the same album usually show a higher level of similarity than other tracks (even by the same artist). To alleviate this issue, some retrieval systems filter results from the same album or even by the same

artist as the seed. Producing a well diversified result set for a given query is thus a common requirement for IR systems.

Familiarity/Popularity vs. Hotness/Trendiness

These four terms or aspects are related to each other. Familiarity or popularity describes how well-known an artist or song is, whereas hotness or trendiness relates to the amount of buzz or attention an artist is currently getting [41]. Popularity has a more positive connotation than the neutral expression of familiarity. However, we will use the terms interchangeably in the remainder of the paper, likewise the terms hotness and trendiness. In terms of temporal aspects, popularity can be seen as a longer lasting property, whereas hotness usually relates to recent appreciation of typically shorter duration, although hot artists might also be very familiar/popular to many people. To give an example, "The Beatles" are certainly popular, whereas "Lady Gaga" currently tends to rank higher on the hotness dimension.

Recentness

This aspect distinguishes recently released pieces from pieces that are older and therefore have a longer (playing) history. In contrast to the aspect of hotness, novelty does not require an artist to be recently popular, just a temporal closeness to the present.

Novelty

This aspect describes whether a music item is novel to the user of the system. If a music recommender keeps on suggesting tracks/artists well-known to the user, he or she will not be satisfied, even if the recommended items are perfectly suited otherwise. Hence, presenting novel recommendations is a vital requirement for a personalized recommender system.

Serendipity

Serendipity is a requirement often mentioned in the context of recommender systems. It means that a user is surprised in a positive way since he discovered an item he did not expect. In the context of music retrieval, we believe that the listener's music preference and taste as well as aspects of artist and song popularity have to be taken into account when we aim at providing serendipitous results. For instance, a fan of medieval folk metal might be rather disappointed and bored if the system recommends the band "Saltatio Mortis", which is very well known for this style of music. In contrast, for a user occasionally enjoying "Metallica" but also "Bob Dylan", the former mentioned band may be a serendipitous recommendation.

Apart from the listener's music preference and taste, a user profile for a serendipitous recommendation algorithm should take into account different categories of users as well as their different cultural backgrounds. For instance, music perception of musicians is likely to be quite dissimilar to that of music experts and editors, which is again different from untrained, passive listeners.

Transparency

For the acceptance of user-aware music retrieval systems it is crucial how the results are presented and explained. The presentation and explanation should be adapted to the users' musical training and preferences. For instance, the system should provide clues about why certain songs have been retrieved: "These two songs are similar because they share the same harmonic progression, the same tempo, are from the same artists, were recorded by the same

producer" or "This song was suggested because you are currently in an aggressive mood while driving your car", or even "This was your favorite song during the Summer you met your future spouse".

4 Adaptive Music Similarity Measures

Users of MIR systems may have a varying (musical) background and experience music in different ways. Consequently, when comparing musical pieces with each other, opinions may diverge. Moreover, different retrieval tasks may also require different views on music similarity. In order to support individual user perspectives and multiple retrieval tasks, an adaptable model of music similarity is required. Often, (dis-)similarity is modeled by a distance measure. Either way, parameters need to be introduced that allow to adapt the measure.

Direct Manipulation (Adaptability)

Depending on how complex the resulting model is, users may be able to manually adjust and tweak the parameters according to their needs. For instance, Baumann et al. [12] describe a joystick interface to control the weights of three similarity facets in a linear combination. From a study with 10 users, it was concluded that users tend to use nearly similar joystick settings throughout different environments for finding a set of similar songs given an anchor song. Though the joystick interface was considered very intuitive by the users, it is unclear whether it may be applied to more than three similarity facets. Similarly, the *E-Mu Jukebox* described by Vignoli et al. [93] allows changing the similarity function that is applied to create a playlist from a seed song. Here, five similarity facets (sound, tempo, mood, genre and year) are visually represented by adapters that can be dragged on a bull's eye. The closer a facet is to the center, the higher is its weight in the similarity computation. Again, a linear weighting scheme is used here. This interface is to some extent scalable with respect to the number of facets but less intuitive. Indeed, a user study with 22 participants showed that the interface is harder to use, but more useful compared to two control systems.

With an increasing number of facets, direct manual manipulation is likely to become more difficult – even for a simple similarity model such as weighted linear combination. Moreover, specific similarity preferences often exist only subconsciously and thus are hard to specify explicitly. Instead of asking the user to explicitly state how he compares music, adaptive MIR systems aim to learn suitable parameter settings from ground truth data (such as expert annotations) or in an interactive way from user feedback.

Query and Relevance Feedback

The content-based MIR system for symbolic music described by Rolland [75] adjusts its similarity model based on feedback received during successive interactions with the user (search sessions). To model the similarity between a transcribed query and a melody, the concept of *pairings* is introduced: A pairing is a part of an alignment (between query and melody) that may comprise several notes and rests. Pairings can be classified into types and for each type, a weight is defined that specifies the importance of the pairing type in the similarity computation. In a ranked list of search results, the user can point out the correct match and optionally some reasonable secondary matches. Given this feedback, the weight for each pairing type is reinforced by a constant update factor if it contributes more to the

similarity in the correct match than in the higher ranked false matches or otherwise decreased respectively. This way, the system can adapt to the user's way of comparing melodies.

The *MUSIPER* system developed by Sotiropoulos et al. [85] constructs music similarity perception models of its users. To this end, users are asked to specify the degree of similarity for retrieved music pieces. The system uses this relevance feedback to train several Radial Basis Function Networks (RBFN) – a special form of neural network – in parallel. Each RBFN represents a different similarity measure based on a different (content-based) feature subset. The model parameters that are adapted during learning are the internal weights of the networks. Finally, the network (and the respective feature subset) which best approximates the similarity ratings specified by the user is selected. The authors report significant improvement of perceived similarity in subsequent music retrievals during an evaluation with 100 participants and argue that the relation between subsets of features and personalized music similarity could be verified.

Collection Clustering

Slaney et al. [84] apply several algorithms based on second-order statistics (whitening, Linear Discriminant Analysis (LDA) [23], Relevant Component Analysis (RCA) [8]) and optimization techniques (Neighborhood Component Analysis (NCA) [26], Large-Margin Nearest Neighbor (LMNN) [94]) to learn Mahalanobis distance metrics for clustering songs by artist, album or blog they appear on. For the optimization, an objective function that mimics the k-nearest neighbor leave-one-out classification error is chosen. Songs are represented as vectors containing various acoustic features. From their experiments, the authors conclude that all algorithms lead to a significant improvement over the baseline. In particular, NCA and RCA showed higher robustness with (artificially generated) noisy features.

The *BeatlesExplorer* [87] (Figure 3, top) is a prototype system for organization and exploration of music collections that adapts to the user's perceived similarity in that it learns weights for different aspects of music similarity. Initially, a growing Self-Organizing Map (SOM) is induced that clusters the music collection. The user has then the possibility to change the location of songs on the map by simple drag-and-drop actions. Each movement of a song causes a weight change in the underlying similarity measure based on a quadratic programming scheme. As a result, the location of other songs may be modified as well. Experiments simulating user interaction with the system show, that during this stepwise adaptation the similarity measure indeed converges to one that captures how the user compares songs.

The *SoniXplorer* [57] shown in Figure 3 (bottom) is another SOM-based system that also adapts a weighted linear combination of basic similarities. Here, the SOM is displayed as video-game-like virtual 3-D landscape accompanied by spatialized playback of songs. Apart from moving songs on the map, the user can raise or lower the terrain to increase or decrease barriers between regions. For the adaptation, a target distance matrix is derived from the arrangement. Then a linear regression learner adapts the weighting accordingly.

Metric Learning with Relative Distance Constraints

In many publications, adapting music similarity is considered as a metric learning problem subject to so-called *relative distance constraints*. A relative distance constraint (s, a, b) demands that the object a is closer to the seed object s than object b, i.e., $d(s, a) < d(s, b)$. Such constraints can be seen as atomic bits of information fed to the adaptation algorithm. They can be derived from a variety of higher-level application-dependent constraints. For

Figure 3 Prototype interfaces for music collection structuring w.r.t. user-adaptive similarity. Top: BeatlesExplorer [87]. Bottom: SoniXplorer [57].

instance, if the user moves a song s from one cluster to a different one in the *BeatlesExplorer* described above, this can be interpreted by the following set of relative distance constraints:

$$d(s, c_t) < d(s, c) \qquad \forall c \in C \setminus \{c_t\}$$

where C is the set of cluster cells of the SOM (each represented by a prototype) and c_t is the target cluster of the user's drag-and-drop action. Bade et al. describe how relative distance constraints can be derived from expert classifications of folk songs [4] or from an existing personal hierarchy of folders with music files [5]. Alternatively, it is also possible to ask the users directly to state the opinion for a triplet of songs as in the bonus round of the *TagATune* game [47]. McFee et al. [64] use artist similarity triples collected in the web survey described by Ellis et al. [24]. They further describe a graph-based technique to detect and remove inconsistencies within sets of constraints such as direct contradictions.

Using relative distance constraints, the task of learning a suitable adaptation of a similarity measure can be formulated as constraint optimization problem. Approaches are manifold and very much depend on the underlying adaptable model of similarity and its parameters. McFee et al. [64] apply a partial order embedding technique that maps artists into multiple non-linear spaces (using different kernel matrices), learns a separate transformation for each kernel, and concatenates the resulting vectors. The Euclidean distance in the resulting embedding space corresponds to the perceived similarity. In further work [62], they use the metric learning to rank (MLR) technique [61] – an extension of the Structural SVM approach [33] – to adapt a Mahalanobis distance according to a ranking loss measure. This approach is also applied by Wolff et al. [97] whose similarity adaptation experiments are based on the *MagnaTagATune* dataset derived from the *TagATune* game [47].

Instead of adapting a Mahalanobis distance, the work of Stober et al. focuses on simpler linear combination models. In [86], they describe various applications and respective adaptation algorithms which they evaluate and compare in [88] also using the *MagnaTagATune* dataset. Their distance model, which is a weighted sum of m facet distances $\delta_{f_1}, \ldots, \delta_{f_m}$, is less expressive because of fewer parameters than the Mahalanobis distance but it can easily be understood and directly manipulated by the user. This design choice specifically addresses the users' desire to remain in control and not to be patronized by an intelligent system that "knows better". Furthermore, this similarity model allows to reformulate the metric learning task as a binary classification problem as described by Cheng et al. [20], which creates the possibility to apply a wide range of sophisticated classification techniques such as SVM. As Figure 4 illustrates, the idea is to rewrite each relative distance constraint $d(s, a) < d(s, b)$ as

$$\sum_{i=1}^{m} w_i(\delta_{f_i}(s, b) - \delta_{f_i}(s, a)) = \sum_{i=1}^{m} w_i x_i = \mathbf{w}^T \mathbf{x} > 0$$

where x_i is the *distance difference* w.r.t. facet f_i. The positive training example $(\mathbf{x}, +1)$ then represents the satisfied constraint whereas the negative example $(-\mathbf{x}, -1)$ represents its violation (i.e., inverting the relation sign). For these training examples, the normal vector of the hyperplane that separates the positive and negative instances contains the adapted facet weights.

5 Novelty and Serendipity in Music Recommendation

The ability to recommend "interesting new music" is considered an important social factor inside communities, especially among groups of young users (and groups of musicians). In

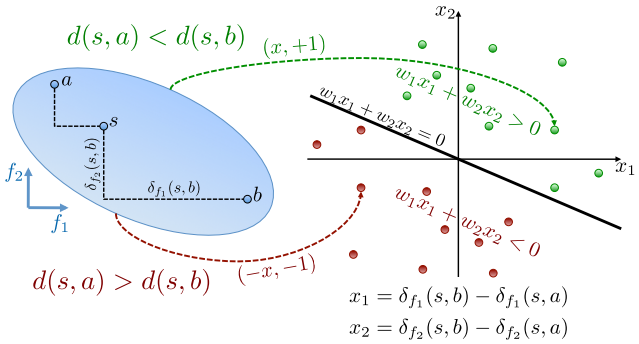


Figure 4 Transformation of a relative distance constraint for linear combination models into two training instances of the corresponding binary classification problem as described in [20].

this context, we use the generic term "music" to address different kinds of recommendations, from individual songs, albums, bands, or sub-genres. A good human recommender takes into account two main components to highlight his role as a music connoisseur:

- He is the first who is aware of music that the others do not know yet, although it is part of their music genre of interest and thus it is likely that sooner or later this music would have been found also without the recommendation.
- He discovers music that might be enjoyed by others, disregarding some aspects of the music content and context that would have suggested the opposite.

In the former case, the emphasis is on the *novelty* of the recommendations, where the role of the human recommender is related to his/her ability to mine music collections and to be at the same time up-to-date with the music market. In the latter case, the emphasis is on *serendipity* because the human recommender can prove his ability to find unexpected relations between music content, pointing towards music that will not be known without his/her recommendation.

Obviously, automatic recommender systems do not have to establish their role inside a community, yet these considerations about what motivates a human recommendation can be a starting point in the development of recommender systems that take into account both novelty and, more important, serendipity. This approach can take advantage of the fact that the user who receives the recommendations can evaluate them also considering how his role in the community will be affected by receiving given recommendations.

From this point of view, the concept of novelty may be extended to include also the process of finding new music. For instance, a user who has in his profile an interest for the recent work of a particular rock band can give a low value to the recommendation of a novel song taken from the band's first recorded album, which can be easily found in any catalogue and a high value to the recommendation of a novel song by another band where some of the musicians he likes appear as guest stars. According to these considerations, the novelty of an item can be measured depending also on the difficulties that a user would encounter to retrieve that particular item in a search session.

Also the concept of serendipity can be partially reconsidered depending on how human recommendations are provided. A central role is played by the fact that the user would not expect to like the recommended music item, because its average characteristics place it far from his listening profile. In order to enjoy the recommended item, the user is required to concentrate on a reduced set – maybe a single aspect – of the music dimensions that

characterize it. For instance, a serendipitous experience for a user with a special interest for classical music for flute is to discover that many background music in movies of the 1970s is played on the flute. Or a serendipitous experience for a user interested in rock music with strong rhythm is to discover Scottish music for drums only.

According to these considerations, serendipity can be related to the ability of selectively suppress some dimensions of music content and context while recommending a list of music items. As a side note, perhaps one of the reasons why pure text-based search systems are still very popular among users of music recommender systems is that they suppress the information which is not explicitly represented in tags and metadata, thus promoting this aspect of serendipity.

6 Conclusions

The contribution of this article is threefold. First, we presented a broad categorization of aspects that influence human music perception, namely computational features related to music content, to music context, and to user context. We briefly reviewed the state-of-the-art in extraction and use of features in each category. Second, we proposed several aspects to take into account when elaborating user-aware music retrieval systems, more precisely, similarity, diversity, familiarity, hotness, recentness, novelty, serendipity, and transparency. Eventually, we thoroughly reported on recent developments in research on adaptive music similarity measures and music recommendation focusing on novelty and serendipity aspects.

We believe that a lot of research is still needed to understand the mechanisms involved in the perception of music similarity according to the three broad categories of aspects. Investigating the relations between computational features and human music perception will eventually pave the way to personalized, user-aware music retrieval systems and therefore is a research endeavor worth pursuing.

References

1 `http://www.amazon.com/music` (access: January 2010).

2 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.

3 Claudio Baccigalupo, Enric Plaza, and Justin Donaldson. Uncovering Affinity of Artists to Multiple Genres from Social Behaviour Data. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, Philadelphia, PA, USA, September 14–18 2008.

4 Korinna Bade, Jörg Garbers, Sebastian Stober, Frans Wiering, and Andreas Nürnberger. Supporting folk-song research by automatic metric learning and ranking. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 741–746, Utrecht, the Netherlands, August 2010.

5 Korinna Bade, Andreas Nürnberger, and Sebastian Stober. Everything in its right place? learning a user's view of a music collection. In *Proceedings of NAG/DAGA 2009, International Conference on Acoustics, Rotterdam*, pages 344–347, 2009.

6 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 2010 Joint Conference on Digital Libraries (JCDL '10)*, pages 149–158, 2010.

7 Linas Baltrunas, Marius Kaminskas, Bernd Ludwig, Omar Moling, Francesco Ricci, Karl-Heinz Lüke, and Roland Schwaiger. InCarMusic: Context-Aware Music Recommendations in a Car. In *International Conference on Electronic Commerce and Web Technologies (EC-Web)*, Toulouse, France, Aug–Sep 2011.

8 Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6(1):937, 2006.

9 Luke Barrington, Reid Oda, and Gert Lanckriet. Smarter than Genius? Human Evaluation of Music Recommender Systems. In *Proc. ISMIR*, pages 357–362, October 2009.

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

11 Mark A. Bartsch and pages=15–18 Gregory H. Wakefield, year=2001. To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics 2001*, October.

12 S. Baumann and J. Halloran. An ecological approach to multimodal subjective music similarity perception. In *Proceedings of 1st Conference on Interdisciplinary Musicology (CIM'04)*, 2004.

13 Dmitry Bogdanov, Martín Haro, Ferdinand Fuhrmann, Anna Xambó, Emilia Gómez, and Perfecto Herrera. A Content-based System for Music Recommendation and Visualization of User Preferences Working on Semantic Notions. In *9th International Workshop on Content-based Multimedia Indexing (CBMI 2011)*, Madrid, Spain, 2011.

14 John S. Breese, David Heckerman, and Carl Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 43–52, San Francisco, CA, USA, 1998. Morgan Kaufmann.

15 Juan José Burred and Alexander Lerch. A Hierarchical Approach to Automatic Musical Genre Classification. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, London, UK, September 8–11 2003.

16 Pedro Cano and Markus Koppenberger. The Emergence of Complex Network Patterns in Music Artist Networks. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR 2004)*, pages 466–469, Barcelona, Spain, October 10–14 2004.

17 Michael A. 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*, 96:668–696, April 2008.

18 Òscar Celma, Miguel Ramírez, and Perfecto Herrera. Foafing the Music: A Music Recommendation System Based on RSS Feeds and User Preferences. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, London, UK, September 11–15 2005.

19 Wei Chai and Barry Vercoe. Using user models in music information retrieval systems. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2000)*, Plymouth, MA, USA, 2000.

20 Weiwei Cheng and Eyke Hüllermeier. Learning similarity functions from qualitative feedback. In *Proceedings of the 9th European Conference on Advances in Case-Based Reasoning (ECCBR'08)*, pages 120–134, 2008.

21 Sally Jo Cunningham, J. Stephen Downie, and David Bainbridge. "The Pain, The Pain": Modelling Music Information Behavior And The Songs We Hate. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 474–477, London, UK, September 11–15 2005.

22 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 2003 Joint Conference on Digital Libraries (JCDL '03), pages 5–16, 2003.

23 J. Duchene and S. Leclercq. An optimal transformation for discriminant and principal component analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 10(6):978–983, 1988.

24 Daniel P.W. Ellis, Brian Whitman, Adam Berenzweig, and Steve Lawrence. The Quest For Ground Truth in Musical Artist Similarity. In *Proceedings of 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, Paris, France, October 13–17 2002.

25 Gijs Geleijnse, Markus Schedl, and Peter Knees. The Quest for Ground Truth in Musical Artist Tagging in the Social Web Era. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, Vienna, Austria, September 23–27 2007.

26 J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

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

28 Fabien Gouyon, Perfecto Herrera, Emilia Gomez, Pedro Cano, Jordi Bonada, Alex Loscos, Xavier Amatriain, and Xavier Serra. Content processing of music audio signals. In Pietro Polotti and Davide Roccheso, editors, *Sound to Sense, Sense to Sound: A State-of-the-art in Sound and Music Computing*, pages 83–160. Logos Verlag, Berlin GmbH, 2008.

29 Fabien Gouyon, François Pachet, and Olivier Delerue. On the Use of Zero-Crossing Rate for an Application of Classification of Percussive Sounds. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx-00)*, Verona, Italy, December 7–9 2000.

30 Sten Govaerts and Erik Duval. A Web-based Approach to Determine the Origin of an Artist. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, Kobe, Japan, October 2009.

31 E. Guaus. *Audio Content Processing for Automatic Music Genre Classification: Descriptors, Databases, and Classifiers*. PhD thesis, Universitat Pompeu Fabra, 2009.

32 Martín Haro, A. Xambó, F. Fuhrmann, D. Bogdanov, E. Gómez, and P. Herrera. The Musical Avatar - A Visualization of Musical Preferences by Means of Audio Content Description. In *5th Audio Mostly Conference: A Conference on Interaction with Sound*, Piteå, Sweden, September 2010.

33 T. Joachims. A support vector method for multivariate performance measures. In *International Conference on Machine Learning (ICML)*, pages 377–384, 2005.

34 Peter Knees, Elias Pampalk, and Gerhard Widmer. Artist Classification with Web-based Data. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR 2004)*, pages 517–524, Barcelona, Spain, October 10–14 2004.

35 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 (SIGIR 2007)*, Amsterdam, the Netherlands, July 23–27 2007.

36 Peter Knees, Markus Schedl, and Gerhard Widmer. Multiple Lyrics Alignment: Automatic Retrieval of Song Lyrics. In *Proceedings of 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 564–569, London, UK, September 11–15 2005.

37 Peter Knees and Gerhard Widmer. Searching for Music Using Natural Language Queries and Relevance Feedback. In *Proceedings of the 5th International Workshop on Adaptive Multimedia Retrieval (AMR'07)*, Paris, France, July 2007.

38 Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Germany, 3rd edition, 2001.

39 Yehuda Koren. Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 426–434, Las Vegas, NV, USA, August 2008.

40 Jan Korst and Gijs Geleijnse. Efficient lyrics retrieval and alignment. In Wim Verhaegh, Emile Aarts, Warner ten Kate, Jan Korst, and Steffen Pauws, editors, *Proceedings of the 3rd Philips Symposium on Intelligent Algorithms (SOIA 2006)*, pages 205–218, Eindhoven, the Netherlands, December 6–7 2006.

41 Paul Lamere. Artist similarity, familiarity and hotness. `http://musicmachinery.com/2009/05/25/artist-similarity-familiarity-and-hotness` (access: September 2011).

42 Audrey Laplante. *Everyday life music information-seeking behaviour of young adults: an exploratory study*. PhD thesis, McGill University, Montréal, Canada, 2008.

43 `http://last.fm` (access: October 2011).

44 C. Laurier and P. Herrera. *Automatic Detection of Emotion in Music: Interaction with Emotionally Sensitive Machines*, chapter 2, pages 9–32. IGI Global, 2009.

45 Cyril Laurier, Jens Grivolla, and Perfecto Herrera. Multimodal Music Mood Classification using Audio and Lyrics. In *Proceedings of the International Conference on Machine Learning and Applications*, San Diego, CA, USA, 2008.

46 E. Law, L. von Ahn, R. Dannenberg, and M. Crawford. Tagatune: A Game for Music and Sound Annotation. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, Vienna, Austria, September 2007.

47 Edith Law and Luis von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings CHI '09*, pages 1197–1206, 2009.

48 Daniel D. Lee and H. Sebastian Seung. Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature*, 401(6755):788–791, 1999.

49 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)*, 61:1025–1045, 2010.

50 Jin Ha Lee. How Similar Is Too Similar?: Exploring Users' Perceptions of Similarity in Playlist Evaluation. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, pages 109–114, Miami, USA, October 2011.

51 Mark Levy and Mark Sandler. A semantic space for music derived from social tags. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, Vienna, Austria, September 2007.

52 Dongge Li, Ishwar K. Sethi, Nevenka Dimitrova, and Tom McGee. Classification of General Audio Data for Content-based Retrieval. *Pattern Recognition Letters*, 22(5):533–544, 2001.

53 Bing Liu. *Web Data Mining – Exploring Hyperlinks, Contents and Usage Data*. Springer, Berlin, Heidelberg, Germany, 2007.

54 Beth Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2000)*, Plymouth, Massachusetts, USA, 2000.

55 Beth Logan, Daniel P.W. Ellis, and Adam Berenzweig. Toward Evaluation Techniques for Music Similarity. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003): Workshop on the Evaluation of Music Information Retrieval Systems*, Toronto, Canada, July–August 2003. ACM Press.

56 Beth Logan, Andrew Kositsky, and Pedro Moreno. Semantic Analysis of Song Lyrics. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2004)*, Taipei, Taiwan, June 27–30 2004.

57 Dominik Lübbers and Matthias Jarke. Adaptive multimodal exploration of music collections. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 195–200, Kobe, Japan, October 2009.

58 `http://mtg.upf.edu/project/musicalavatar` (access: October 2011).

59 Michael I. Mandel and Daniel P.W. Ellis. A Web-based Game for Collecting Music Metadata. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, Vienna, Austria, September 2007.

60 Rudolf Mayer, Robert Neumayer, and Andreas Rauber. Rhyme and Style Features for Musical Genre Classification by Song Lyrics. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, 2008.

61 B. McFee and G. R. G. Lanckriet. Metric learning to rank. In *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*, 2010.

62 Brian McFee, Luke Barrington, and G.R.G. Lanckriet. Learning similarity from collaborative filters. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 345–350, Utrecht, the Netherlands, August 2010.

63 Brian McFee and Gert Lanckriet. Heterogeneous Embedding for Subjective Artist Similarity. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, Kobe, Japan, October 2009.

64 Brian McFee and Gert Lanckriet. Heterogeneous embedding for subjective artist similarity. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 513–518, Kobe, Japan, October 2009.

65 M. Müller, D.P.W. Ellis, A. Klapuri, and G. Richard. Signal processing for music analysis. *IEEE Journal on Selected Topics in Signal Processing*, 5(6):1088–1110, October 2011.

66 Andreas Nürnberger and Marcin Detyniecki. Weighted Self-Organizing Maps: Incorporating User Feedback. In Okyay Kaynak and Erkki Oja, editors, *Proceedings of the Joined 13th International Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP 2003)*, pages 883–890. Springer-Verlag, 2003.

67 François Pachet, Gert Westerman, and Damien Laigre. Musical Data Mining for Electronic Music Distribution. In *Proceedings of the 1st International Conference on Web Delivering of Music (WEDELMUSIC 2001)*, Florence, Italy, November 23–24 2001.

68 Elias Pampalk. *Computational Models of Music Similarity and their Application to Music Information Retrieval*. PhD thesis, Vienna University of Technology, March 2006.

69 Elias Pampalk, Andreas Rauber, and Dieter Merkl. Content-based Organization and Visualization of Music Archives. In *Proceedings of the 10th ACM International Conference on Multimedia (MM 2002)*, pages 570–579, Juan les Pins, France, December 1–6 2002.

70 Tim Pohle. *Automatic Characterization of Music for Intuitive Retrieval*. PhD thesis, Johannes Kepler University Linz, Linz, Austria, 2009.

71 Tim Pohle, Peter Knees, Markus Schedl, Elias Pampalk, and Gerhard Widmer. "Reinventing the Wheel": A Novel Approach to Music Player Interfaces. *IEEE Transactions on Multimedia*, 9:567–575, 2007.

72 Tim Pohle, Peter Knees, Markus Schedl, and Gerhard Widmer. Building an Interactive Next-Generation Artist Recommender Based on Automatically Derived High-Level Concepts. In *Proceedings of the 5th International Workshop on Content-Based Multimedia Indexing (CBMI'07)*, Bordeaux, France, June 2007.

73 G. Poliner, D. Ellis, A. Ehmann, E. Gómez, S. Streich, and B. Ong. Melody transcription from music audio approaches and evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 15:1247–1256, 2007.

74 Joseph J. Rocchio. Relevance Feedback in Information Retrieval. In Gerard Salton, editor, *The SMART Retrieval System - Experiments in Automatic Document Processing*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall, 1971.

75 P.Y. Rolland. Adaptive user modeling in a content-based music retrieval system. In *Proceedings of the 2nd International Conference on Music Information Retrieval (ISMIR 2001)*, Bloomington, Indiana, USA, October 2001.

76 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.

77 Markus Schedl, Peter Knees, and Gerhard Widmer. A Web-Based Approach to Assessing Artist Similarity using Co-Occurrences. In *Proceedings of the 4th International Workshop on Content-Based Multimedia Indexing (CBMI 2005)*, Riga, Latvia, June 21–23 2005.

78 Markus Schedl, Elias Pampalk, and Gerhard Widmer. Intelligent Structuring and Exploration of Digital Music Collections. *e&i - Elektrotechnik und Informationstechnik*, 122(7–8):232–237, July–August 2005.

79 Markus Schedl, Tim Pohle, Peter Knees, and Gerhard Widmer. Assigning and Visualizing Music Genres by Web-based Co-Occurrence Analysis. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, Victoria, Canada, October 8–12 2006.

80 Markus Schedl, Cornelia Schiketanz, and Klaus Seyerlehner. Country of Origin Determination via Web Mining Techniques. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2010): 2nd International Workshop on Advances in Music Information Research (AdMIRe 2010)*, Singapore, July 19–23 2010.

81 Markus Schedl, Klaus Seyerlehner, Dominik Schnitzer, Gerhard Widmer, and Cornelia Schiketanz. Three Web-based Heuristics to Determine a Person's or Institution's Country of Origin. In *Proceedings of the 33th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, Geneva, Switzerland, July 19–23 2010.

82 Eric Scheirer and Malcolm Slaney. Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1997)*, pages 1331–1334, Munich, Germany, April 21–24 1997.

83 Yuval Shavitt and Udi Weinsberg. Songs Clustering Using Peer-to-Peer Co-occurrences. In *Proceedings of the IEEE International Symposium on Multimedia (ISM2009): International Workshop on Advances in Music Information Research (AdMIRe 2009)*, San Diego, CA, USA, December 16 2009.

84 Malcolm Slaney, Kilian Q. Weinberger, and William White. Learning a metric for music similarity. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, pages 313–318, Philadelphia, PA, USA, September 2008.

85 Dionysios N. Sotiropoulos, Aristomenis S. Lampropoulos, and George A. Tsihrintzis. Musiper: a system for modeling music similarity perception based on objective feature subset selection. *User Modeling and User-Adapted Interaction*, 18(4):315–348, 2008.

86 Sebastian Stober. Adaptive distance measures for exploration and structuring of music collections. In *Proceedings of AES 42nd Conference on Semantic Audio*, 2011.

87 Sebastian Stober and Andreas Nürnberger. Towards user-adaptive structuring and organization of music collections. In *Proceedings of the 6th international workshop on Adaptive Multimedia Retrieval (AMR'08)*, 2008.

88 Sebastian Stober and Andreas Nürnberger. An experimental comparison of similarity adaptation approaches. In *Proceedings of 9th International Workshop on Adaptive Multimedia Retrieval (AMR'11)*, 2011.

89 Sebastian Stober, Matthias Steinbrecher, and Andreas Nüurnberger. A Survey on the Acceptance of Listening Context Logging for MIR Applications. In *Proceedings of 3rd Work-*

shop on Learning the Semantics of Audio Signals (LSAS 2009), Graz, Austria, December 2009.

90 Sebastian Streich. *Music Complexity: A Multi-faceted Description of Audio Content*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2007.

91 D. Turnbull, R. Liu, L. Barrington, and G. Lanckriet. A Game-based Approach for Collecting Semantic Annotations of Music. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, Vienna, Austria, September 2007.

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

93 Fabio Vignoli and Steffen Pauws. A music retrieval system based on user driven similarity and its evaluation. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 272–279, London, UK, September 2005.

94 K.Q. Weinberger, J. Blitzer, and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

95 Brian Whitman, Gary Flake, and Steve Lawrence. Artist Detection in Mmusic with Minnowmatch. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, pages 559–568, Falmouth, MA, USA, September 10-12 2001.

96 Brian Whitman and Steve Lawrence. Inferring Descriptions and Similarity for Music from Community Metadata. In *Proceedings of the 2002 International Computer Music Conference (ICMC 2002)*, pages 591–598, Göteborg, Sweden, September 16–21 2002.

97 Daniel Wolff and Tillman Weyde. Combining sources of description for approximating music similarity ratings. In *Proceedings of the 9th International Workshop on Adaptive Multimedia Retrieval (AMR'11)*, 2011.

98 Wei Xu, Xin Liu, and Yihong Gong. Document Clustering Based on Non-negative Matrix Factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, pages 267–273, Toronto, Canada, July 28–August 1 2003. ACM Press.

99 Gui-Rong Xue, Jie Han, Yong Yu, and Qiang Yang. User Language Model for Collaborative Personalized Search. *ACM Transactions on Information Systems*, 27(2), February 2009.

100 Bingjun Zhang, Jialie Shen, Qiaoliang Xiang, and Ye Wang. CompositeMap: A Novel Framework for Music Similarity Measure. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, pages 403–410, New York, NY, USA, 2009. ACM.

101 Bingjun Zhang, Qiaoliang Xiang, Ye Wang, and Jialie Shen. CompositeMap: A Novel Music Similarity Measure for Personalized Multimodal Music Search. In *MM '09: Proceedings of the seventeen ACM international conference on Multimedia*, pages 973–974, New York, NY, USA, 2009. ACM.

Audio Content-Based Music Retrieval

Peter Grosche[*1], Meinard Müller[*1], and Joan Serrà[†2]

1 **Saarland University and MPI Informatik**
Campus E1-4, 66123 Saarbrücken, Germany
pgrosche@mpi-inf.mpg.de, meinard@mpi-inf.mpg.de

2 **Artificial Intelligence Research Institute (IIIA-CSIC)**
Campus UAB s/n, 08193 Bellaterra, Barcelona, Spain
jserra@iiia.csic.es

Abstract

The rapidly growing corpus of digital audio material requires novel retrieval strategies for exploring large music collections. Traditional retrieval strategies rely on metadata that describe the actual audio content in words. In the case that such textual descriptions are not available, one requires content-based retrieval strategies which only utilize the raw audio material. In this contribution, we discuss content-based retrieval strategies that follow the query-by-example paradigm: given an audio query, the task is to retrieve all documents that are somehow similar or related to the query from a music collection. Such strategies can be loosely classified according to their *specificity*, which refers to the degree of similarity between the query and the database documents. Here, high specificity refers to a strict notion of similarity, whereas low specificity to a rather vague one. Furthermore, we introduce a second classification principle based on *granularity*, where one distinguishes between fragment-level and document-level retrieval. Using a classification scheme based on specificity and granularity, we identify various classes of retrieval scenarios, which comprise *audio identification*, *audio matching*, and *version identification*. For these three important classes, we give an overview of representative state-of-the-art approaches, which also illustrate the sometimes subtle but crucial differences between the retrieval scenarios. Finally, we give an outlook on a user-oriented retrieval system, which combines the various retrieval strategies in a unified framework.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems, I.5 Pattern Recognition

Keywords and phrases music retrieval, content-based, query-by-example, audio identification, audio matching, cover song identification

Digital Object Identifier 10.4230/DFU.Vol3.11041.157

1 Introduction

The way music is stored, accessed, distributed, and consumed underwent a radical change in the last decades. Nowadays, large collections containing millions of digital music documents are accessible from anywhere around the world. Such a tremendous amount of readily available music requires retrieval strategies that allow users to explore large music collections in a convenient and enjoyable way. Most audio search engines rely on metadata and textual

* The authors are funded by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI). Meinard Müller is now with Bonn University, Department of Computer Science III, Germany.
† Funded by Consejo Superior de Investigaciones Científicas (JAEDOC069/2010) and Generalitat de Catalunya (2009-SGR-1434).

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 157–174

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

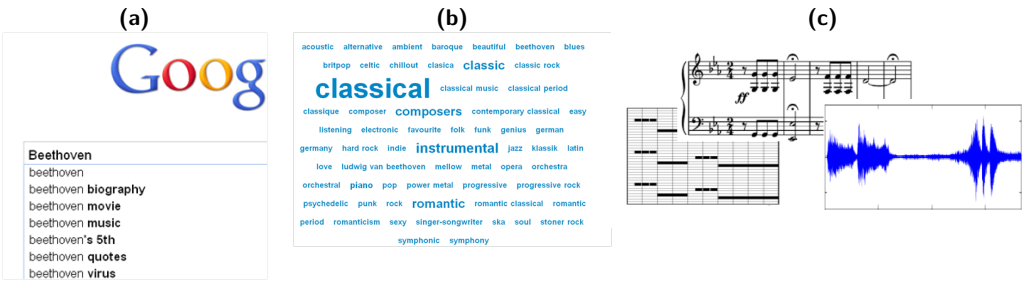


Figure 1 Illustration of retrieval concepts. **(a)** Traditional retrieval using textual metadata (e. g., artist, title) and a web search engine.[1] **(b)** Retrieval based on rich and expressive metadata given by tags.[2] **(c)** Content-based retrieval using audio, MIDI, or score information.

annotations of the actual audio content [11]. Editorial metadata typically include descriptions of the artist, title, or other release information. The drawback of a retrieval solely based on editorial metadata is that the user needs to have a relatively clear idea of what he or she is looking for. Typical query terms may be a title such as "Act naturally" when searching the song by The Beatles or a composer's name such as "Beethoven" (see Figure 1a). In other words, traditional editorial metadata only allow to search for already known content. To overcome these limitations, editorial metadata has been more and more complemented by general and expressive annotations (so called *tags*) of the actual musical content [5, 25, 49]. Typically, tags give descriptions of the musical style or genre of a recording, but may also include information about the mood, the musical key, or the tempo [31, 48]. In particular, tags form the basis for music recommendation and navigation systems that make the audio content accessible even when users are not looking for a specific song or artist but for music that exhibits certain musical properties [49]. The generation of such annotations of audio content, however, is typically a labor intensive and time-consuming process [11, 48]. Furthermore, often musical expert knowledge is required for creating reliable, consistent, and musically meaningful annotations. To avoid this tedious process, recent attempts aim at substituting expert-generated tags by user-generated tags [48]. However, such tags tend to be less accurate, subjective, and rather noisy. In other words, they exhibit a high degree of variability between users. Crowd (or social) tagging, one popular strategy in this context, employs voting and filtering strategies based on large social networks of users for "cleaning" the tags [31]. Relying on the "wisdom of the crowd" rather than the "power of the few" [27], tags assigned by many users are considered more reliable than tags assigned by only a few users. Figure 1b shows the Last.fm[2] *tag cloud* for "Beethoven". Here, the font size reflects the frequency of the individual tags. One major drawback of this approach is that it relies on a large crowd of users for creating reliable annotations [31]. While mainstream pop/rock music is typically covered by such annotations, less popular genres are often scarcely tagged. This phenomenon is also known as the "long-tail" problem [12, 48]. To overcome these problems, *content-based retrieval* strategies have great potential as they do not rely on any manually created metadata but are exclusively based on the audio content and cover the entire audio material in an objective and reproducible way [11]. One possible approach is to employ automated procedures for tagging music, such as automatic genre recognition, mood recognition, or tempo estimation [4, 49]. The major drawback of these learning-based

[1] `www.google.com` (accessed Dec. 18, 2011)
[2] `www.last.fm` (accessed Dec. 18, 2011)

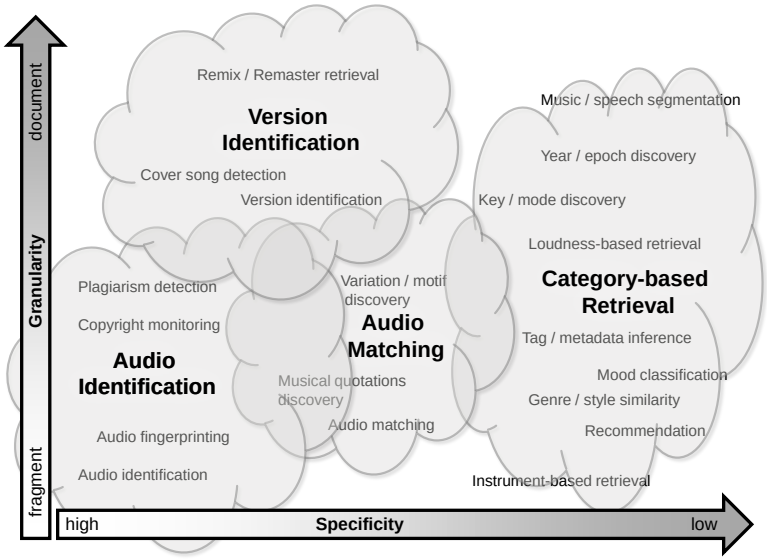


Figure 2 Specificity/granularity pane showing the various facets of content-based music retrieval.

strategies is the requirement of large corpora of tagged music examples as training material and the limitation to queries in textual form. Furthermore, the quality of the tags generated by state-of-the-art procedures does not reach the quality of human generated tags [49].

In this contribution, we present and discuss various retrieval strategies based on audio content that follow the query-by-example paradigm: given an audio recording or a fragment of it (used as query or example), the task is to automatically retrieve documents from a given music collection containing parts or aspects that are similar to it. As a result, retrieval systems following this paradigm do not require any textual descriptions. However, the notion of similarity used to compare different audio recordings (or fragments) is of crucial importance and largely depends on the respective application as well as the user requirements.

Many different audio content-based retrieval systems have been proposed, following different strategies and aiming at different application scenarios. Generally, such retrieval systems can be characterized by various aspects such as the notion of similarity, the underlying matching principles, or the query format. Following and extending the concept introduced in [11], we consider the following two aspects: *specificity* and *granularity*, see Figure 2. The *specificity* of a retrieval system refers to the degree of similarity between the query and the database documents to be retrieved. High-specific retrieval systems return exact copies of the query (in other words, they *identify* the query or occurrences of the query within database documents), whereas low-specific retrieval systems return vague matches that are similar with respect to some musical properties. As in [11], different content-based music retrieval scenarios can be arranged along a specificity axis as shown in Figure 2 (horizontally). We extend this classification scheme by introducing a second aspect, the *granularity* (or temporal scope) of a retrieval scenario. In *fragment-level* retrieval scenarios, the query consists of a short fragment of an audio recording, and the goal is to retrieve all musically related fragments that are contained in the documents of a given music collection. Typically, such fragments may cover only a few seconds of audio content or may correspond to a motif, a theme, or a musical part of a recording. In contrast, in *document-level* retrieval, the query reflects characteristics of an entire document and is compared with entire documents of the database.

Here, the notion of similarity typically is rather coarse and the used features capture global statistics of an entire recording. In this context, one has to distinguish between some kind of internal and some kind of external granularity of the retrieval tasks. In our classification scheme, we use the term fragment-level when a fragment-based similarity measure is used to compare fragments of audio recordings (internal), even though entire documents are returned as matches (external). Using such a classification allows for extending the specificity axis to a specificity/granularity pane as shown in Figure 2. In particular, we have identified four different groups of retrieval scenarios corresponding to the four clouds in Figure 2. Each of the clouds, in turn, encloses a number of different retrieval scenarios. Obviously, the clouds are not strictly separated but blend into each other. Even though this taxonomy is rather vague and sometimes questionable, it gives an intuitive overview of the various retrieval paradigms while illustrating their subtle but crucial differences.

An example of a high-specific fragment-level retrieval task is *audio identification* (sometimes also referred to as *audio fingerprinting* [8]). Given a small audio fragment as query, the task of audio identification consists in identifying the particular audio recording that is the source of the fragment [1]. Nowadays, audio identification is widely used in commercial systems such as Shazam.[3] Typically, the query fragment is exposed to signal distortions on the transmission channel [8, 29]. Recent identification algorithms exhibit a high degree of robustness against noise, MP3 compression artifacts, uniform temporal distortions, or interferences of multiple signals [16, 22]. The high specificity of this retrieval task goes along with a notion of similarity that is very close to the identity. To make this point clearer, we distinguish between a piece of music (in an abstract sense) and a specific performance of this piece. In particular for Western classical music, there typically exist a large number of different recordings of the same piece of music performed by different musicians. Given a query fragment, e. g., taken from a Bernstein recording of Beethoven's Symphony No. 5, audio fingerprinting systems are not capable of retrieving, e. g., a Karajan recording of the same piece. Likewise, given a query fragment from a live performance of "Act naturally" by The Beatles, the original studio recording of this song may not be found. The reason for this is that existing fingerprinting algorithms are not designed to deal with strong non-linear temporal distortions or with other musically motivated variations that affect, for example, the tempo or the instrumentation.

At a lower specificity level, the goal of fragment-based *audio matching* is to retrieve all audio fragments that musically correspond to a query fragment from all audio documents contained in a given database [28, 37]. In this scenario, one explicitly allows semantically motivated variations as they typically occur in different performances and arrangements of a piece of music. These variations include significant non-linear global and local differences in tempo, articulation, and phrasing as well as differences in executing note groups such as grace notes, trills, or arpeggios. Furthermore, one has to deal with considerable dynamical and spectral variations, which result from differences in instrumentation and loudness.

One instance of document-level retrieval at a similar specificity level as audio matching is the task of *version identification*. Here, the goal is to identify different versions of the same piece of music within a database [42]. In this scenario, one not only deals with changes in instrumentation, tempo, and tonality, but also with more extreme variations concerning the musical structure, key, or melody, as typically occurring in remixes and cover songs. This requires document-level similarity measures to globally compare entire documents.

Finally, there are a number of even less specific document-level retrieval tasks which

[3] `www.shazam.com` (accessed Dec. 18, 2011)

can be grouped under the term *category-based retrieval*. This term encompasses retrieval of documents whose relationship can be described by cultural or musicological categories. Typical categories are genre [50], rhythm styles [19, 41], or mood and emotions [26, 47, 53] and can be used in fragment as well as document-level retrieval tasks. Music recommendation or general music similarity assessments [7, 54] can be seen as further document-level retrieval tasks of low specificity.

In the following, we elaborate the aspects of specificity and granularity by means of representative state-of-the-art content-based retrieval approaches. In particular, we highlight characteristics and differences in requirements when designing and implementing systems for audio identification, audio matching, and version identification. Furthermore, we address efficiency and scalability issues. We start with discussing high-specific audio fingerprinting (Section 2), continue with mid-specific audio matching (Section 3), and then discuss version identification (Section 4). In Section 5, we discuss open problems in the field of content-based retrieval and give an outlook on future directions.

2 Audio Identification

Of all content-based music retrieval tasks, audio identification has received most interest and is now widely used in commercial applications. In the identification process, the audio material is compared by means of so-called *audio fingerprints*, which are compact content-based signatures of audio recordings [8]. In real-world applications, these fingerprints need to fulfill certain requirements. First of all, the fingerprints should capture highly specific characteristics so that a short audio fragment suffices to reliably identify the corresponding recording and distinguish it from millions of other songs. However, in real-world scenarios, audio signals are exposed to distortions on the transmission channel. In particular, the signal is likely to be affected by noise, artifacts from lossy audio compression, pitch shifting, time scaling, equalization, or dynamics compression. For a reliable identification, fingerprints have to show a significant degree of robustness against such distortions. Furthermore, scalability is an important issue for all content-based retrieval applications. A reliable audio identification system needs to capture the entire digital music catalog, which is further growing every day. In addition, to minimize storage requirements and transmission delays, fingerprints should be compact and efficiently computable [8]. Most importantly, this also requires efficient retrieval strategies to facilitate very fast database look-ups. These requirements are crucial for the design of large-scale audio identification systems. To satisfy all these requirements, however, one typically has to face a trade-off between contradicting principles.

There are various ways to design and compute fingerprints. One group of fingerprints consist of short sequences of frame-based feature vectors such as Mel-Frequency Cepstral Coefficients (MFCC) [9], Bark-scale spectrograms [22, 23], or a set of low-level descriptors [1]. For such representations, vector quantization [1] or thresholding [22] techniques, or temporal statistics [38] are needed for obtaining the required robustness. Another group of fingerprints consist of a sparse set of characteristic points such as spectral peaks [14, 52] or characteristic wavelet coefficients [24]. As an example, we now describe the peak-based fingerprints suggested by Wang [52], which are now commercially used in the Shazam music identification service[4].

The Shazam system provides a smartphone application that allows users to record a short audio fragment of an unknown song using the built-in microphone. The application

[4] `www.shazam.com` (accessed Dec. 18, 2011)

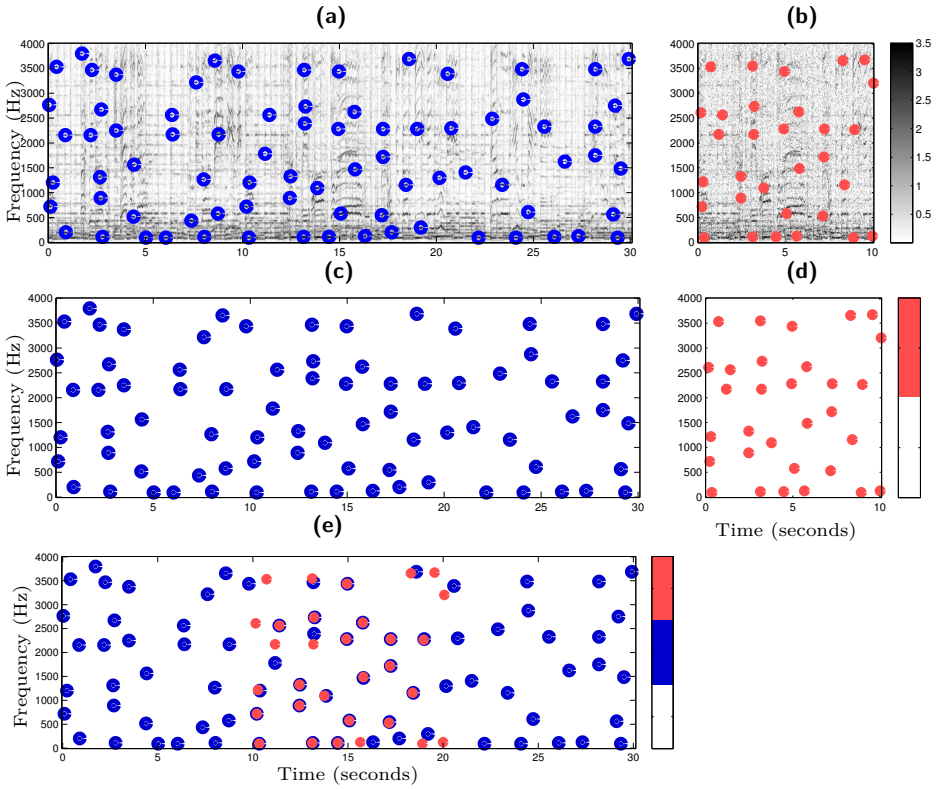


Figure 3 Illustration of the Shazam audio identification system using a recording of "Act naturally" by The Beatles as example. **(a)** Database document with extracted peak fingerprints. **(b)** Query fragment (10 seconds) with extracted peak fingerprints. **(c)** Constellation map of database document. **(d)** Constellation map of query document. **(e)** Superposition of the database fingerprints and time-shifted query fingerprints.

then derives the audio fingerprints which are sent to a server that performs the database look-up. The retrieval result is returned to the application and presented to the user together with additional information about the identified song. In this approach, one first computes a spectrogram from an audio recording using a short-time Fourier transform. Then, one applies a peak-picking strategy that extracts local maxima in the magnitude spectrogram: time-frequency points that are locally predominant. Figure 3 illustrates the basic retrieval concept of the Shazam system using a recording of "Act naturally" by The Beatles. Figure 3a and Figure 3b show the spectrogram for an example database document (30 seconds of the recording) and a query fragment (10 seconds), respectively. The extracted peaks are superimposed to the spectrograms. The peak-picking step reduces the complex spectrogram to a "constellation map", a low-dimensional sparse representation of the original signal by means of a small set of time-frequency points, see Figure 3c and Figure 3d. According to [52], the peaks are highly characteristic, reproducible, and robust against many, even significant distortions of the signal. Note that a peak is only defined by its time and frequency values, whereas magnitude values are no longer considered.

The general database look-up strategy works as follows. Given the constellation maps for a query fragment and all database documents, one locally compares the query fragment to all database fragments of the same size. More precisely, one counts matching peaks, i. e., peaks that occur in both constellation maps. A high count indicates that the corresponding database fragment is likely to be a correct hit. This procedure is illustrated in Figure 3e,

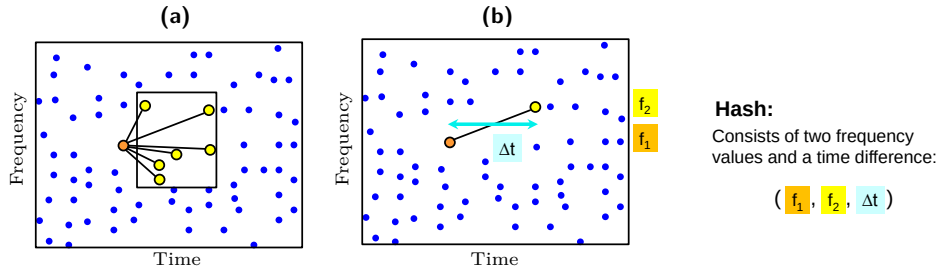


Figure 4 Illustration of the peak pairing strategy of the Shazam algorithm. **(a)** Anchor peak and assigned target zone. **(b)** Pairing of anchor peak and target peaks to form hash values.

showing the superposition of the database fingerprints and time-shifted query fingerprints. Both constellation maps show a high consistency (many red and blue points coincide) at a fragment of the database document starting at time position 10 seconds, which indicates a hit. However, note that not all query and database peaks coincide. This is because the query was exposed to signal distortions on the transmission channel (in this example additive white noise). Even under severe distortions of the query, there still is a high number of coinciding peaks thus showing the robustness of these fingerprints.

Obviously, such an exhaustive search strategy is not feasible for a large database as the run-time linearly depends on the number and sizes of the documents. For the constellation maps, as proposed in [29], one tries to efficiently reduce the retrieval time using indexing techniques—very fast operations with a sub-linear run-time. However, directly using the peaks as hash values is not possible as the temporal component is not translation-invariant and the frequency component alone does not have the required specificity. In [52], a strategy is proposed, where one considers pairs of peaks. Here, one first fixes a peak to serve as "anchor peak" and then assigns a "target zone" as indicated in Figure 4a. Then, pairs are formed of the anchor and each peak in the target zone, and a hash value is obtained for each pair of peaks as a combination of both frequency values and the time difference between the peaks as indicated in Figure 4b. Using every peak as anchor peak, the number of items to be indexed increases by a factor that depends on the number of peaks in the target zone. This combinatorial hashing strategy has three advantages. Firstly, the resulting fingerprints show a higher specificity than single peaks, leading to an acceleration of the retrieval as fewer exact hits are found. Secondly, the fingerprints are translation-invariant as no absolute timing information is captured. Thirdly, the combinatorial multiplication of the number of fingerprints introduced by considering pairs of peaks as well as the local nature of the peak pairs increases the robustness to signal degradations.

The Shazam audio identification system facilitates a high identification rate, while scaling to large databases. One weakness of this algorithm is that it can not handle time scale modifications of the audio as frequently occurring in the context of broadcasting monitoring. The reason for this is that time scale modifications (also leading to frequency shifts) of the query fragment completely change the hash values. Extensions of the original algorithms dealing with this issue are presented in [14, 51].

3 Audio Matching

The problem of audio identification can be regarded as largely solved even for large scale music collections. Less specific retrieval tasks, however, are still mostly unsolved. In this

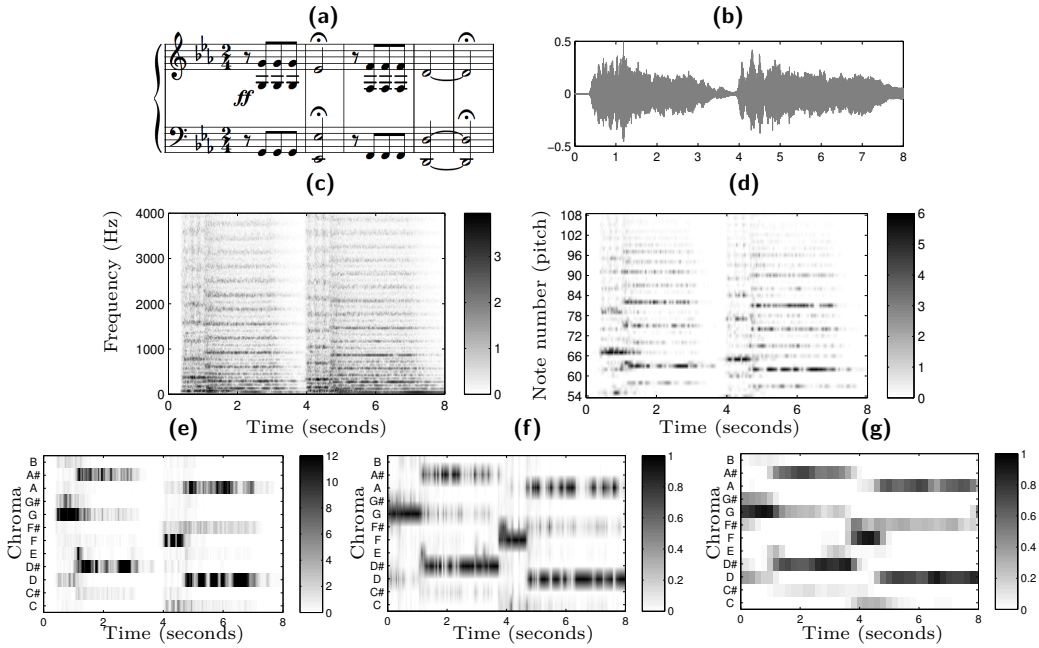


Figure 5 Illustration of various feature representations for the beginning of Beethoven's Opus 67 (Symphony No. 5) in a Bernstein interpretation. **(a)** Score of the excerpt. **(b)** Waveform. **(c)** Spectrogram with linear frequency axis. **(d)** Spectrogram with frequency axis corresponding to musical pitches. **(e)** Chroma features. **(f)** Normalized chroma features. **(g)** Smoothed version of chroma features, see also [36].

section, we highlight the difference between high-specific audio identification and mid-specific audio matching while presenting strategies to cope with musically motivated variations. In particular, we introduce chroma-based audio features [2, 17, 34] and sketch distance measures that can deal with local tempo distortions. Finally, we indicate how the matching procedure may be extended using indexing methods to scale to large datasets [10, 28].

For the audio matching task, suitable descriptors are required to capture characteristics of the underlying piece of music, while being invariant to properties of a particular recording. Chroma-based audio features [2, 34], sometimes also referred to as pitch class profiles [17], are a well-established tool for analyzing Western tonal music and have turned out to be a suitable mid-level representation in the retrieval context [10, 28, 37, 34]. Assuming the equal-tempered scale, the chroma attributes correspond to the set $\{\mathrm{C}, \mathrm{C}^\sharp, \mathrm{D}, \ldots, \mathrm{B}\}$ that consists of the twelve pitch spelling attributes as used in Western music notation. Capturing energy distributions in the twelve pitch classes, chroma-based audio features closely correlate to the harmonic progression of the underlying piece of music. This is the reason why basically every matching procedure relies on some type of chroma feature.

There are many ways for computing chroma features. For example, the decomposition of an audio signal into a chroma representation (or chromagram) may be performed either by using short-time Fourier transforms in combination with binning strategies [17] or by employing suitable multirate filter banks [34, 36]. Figure 5 illustrates the computation of chroma features for a recording of the first five measures of Beethoven's Symphony No. 5 in a Bernstein interpretation. The main idea is that the fine-grained (and highly specific) signal representation as given by a spectrogram (Figure 5c) is coarsened in a musically meaningful way. Here, one adapts the frequency axis to represent the semitones of the equal tempered scale (Figure 5d). The resulting representation captures musically relevant pitch

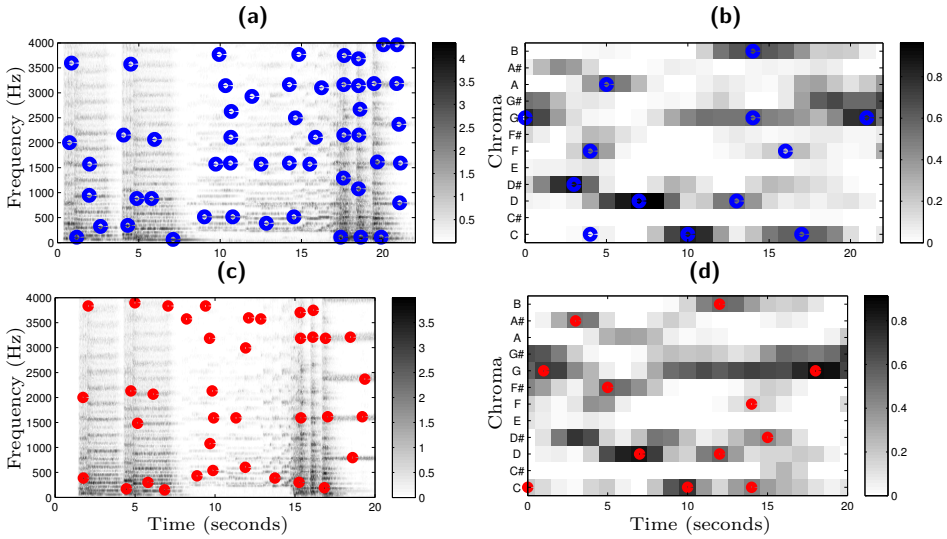


Figure 6 Different representations and peak fingerprints extracted for recordings of the first 21 measures of Beethoven's Symphony No. 5. **(a)** Spectrogram-based peaks for a Bernstein recording. **(b)** Chromagram-based peaks for a Bernstein recording. **(c)** Spectrogram-based peaks for a Karajan recording. **(d)** Chromagram-based peaks for a Karajan recording.

information of the underlying music piece, while being significantly more robust against spectral distortions than the original spectrogram. To obtain chroma features, pitches differing by octaves are summed up to yield a single value for each pitch class, see Figure 5e. The resulting chroma features show increased robustness against changes in timbre, as typically resulting from different instrumentations.

The degree of robustness of the chroma features against musically motivated variations can be further increased by using suitable post-processing steps. See [36] for some chroma variants.[5] For example, normalizing the chroma vectors (Figure 5f) makes the features invariant to changes in loudness or dynamics. Furthermore, applying a temporal smoothing and downsampling step (see Figure 5g) may significantly increase robustness against local temporal variations that typically occur as a result of local tempo changes or differences in phrasing and articulation. There are many more variants of chroma features comprising various processing steps. For example, applying logarithmic compression or whitening procedures enhances small yet perceptually relevant spectral components and the robustness to timbre [33, 35]. A peak picking of spectrum's local maxima can enhance harmonics while suppressing noise-like components [17, 13]. Furthermore, generalized chroma representations with 24 or 36 bins (instead of the usual 12 bins) allow for dealing with differences in tuning [17]. Such variations in the feature extraction pipeline have a large influence and the resulting chroma features can behave quite differently in the subsequent analysis task.

Figure 6 shows spectrograms and chroma features for two different interpretations (by Bernstein and Karajan) of Beethoven's Symphony No. 5. Obviously, the chroma features exhibit a much higher similarity than the spectrograms, revealing the increased robustness against musical variations. The fine-grained spectrograms, however, reveal characteristics of the individual interpretations. To further illustrate this, Figure 6 also shows fingerprint peaks

[5] MATLAB implementations for some chroma variants are supplied by the Chroma Toolbox: `www.mpi-inf.mpg.de/resources/MIR/chromatoolbox` (accessed Dec. 18, 2011)

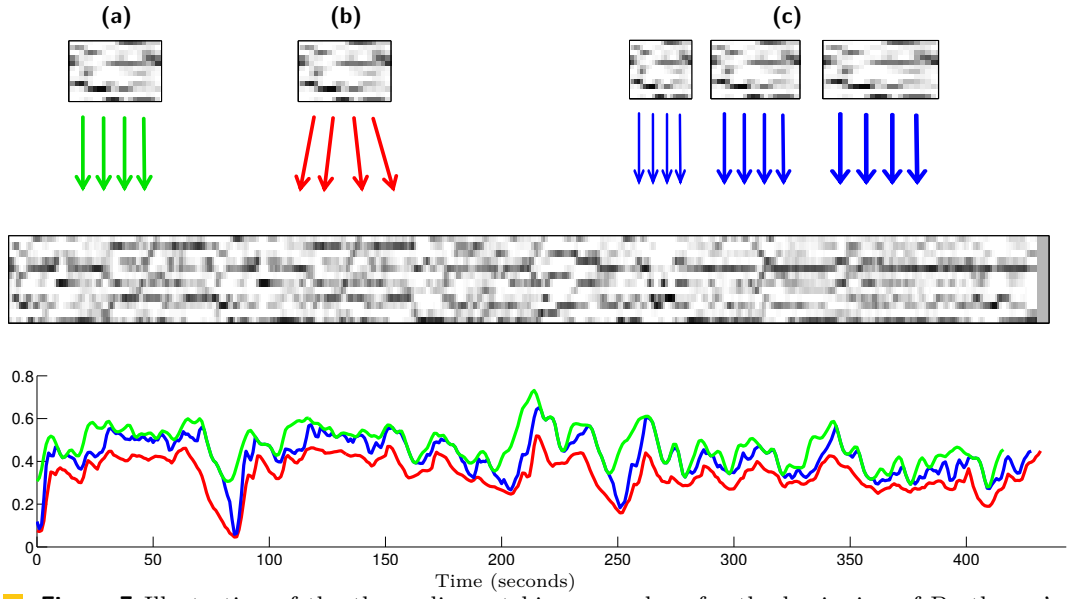


Figure 7 Illustration of the the audio matching procedure for the beginning of Beethoven's Opus 67 (Symphony No. 5) using a query fragment corresponding to the first 22 seconds (measures 1-21) of a Bernstein interpretation and a database consisting of an entire recording of a Karajan interpretation. Three different strategies are shown leading to three different matching curves. **(a)** Strict subsequence matching. **(b)** DTW-based matching. **(c)** Multiple query scaling strategy.

for all representations. As expected, the spectrogram peaks are very inconsistent for the different interpretations. The chromagram peaks, however, show at least some consistencies, indicating that fingerprinting techniques could also be applicable for audio matching [6]. In practice, however, the fragile peak picking step on the basis of the rather coarse chroma features may not lead to robust results. Furthermore, one has to find a technique to deal with the local and global tempo differences between the interpretations. See [21] for a detailed investigation of this approach.

Instead of using sparse peak representations, one typically employs a subsequence search, which is directly performed on the chroma features. Here, a query chromagram is compared with all subsequences of database chromagrams. As a result one obtains a matching curve as shown in Figure 7, where a small value indicates that the subsequence of the database starting at this position is similar to the query sequence. Then the best match is the minimum of the matching curve. In this context, one typically applies distance measures that can deal with tempo differences between the versions, such as edit distances [3], dynamic time warping (DTW) [34, 37], or the Smith-Waterman algorithm [43]. An alternative approach is to linearly scale the query to simulate different tempi and then to minimize over the distances obtained for all scaled variants [28]. Figure 7 shows three different matching curves which are obtained using strict subsequence matching, DTW, and a multiple query strategy.

To speed up such exhaustive matching procedures, one requires methods that allow for efficiently detecting *near* neighbors rather than exact matches. A first approach in this direction uses inverted file indexing [28] and depends on a suitable codebook consisting of a finite set of characteristic chroma vectors. Such a codebook can be obtained in an unsupervised way using vector quantization or in a supervised way exploiting musical knowledge about chords. The codebook then allows for classifying the chroma vectors of the database and to index the vectors according to the assigned codebook vector. This results in

an inverted list for each codebook vector. Then, an exact search can be performed efficiently by intersecting suitable inverted lists. However, the performance of the exact search using quantized chroma vectors greatly depends on the codebook. This requires fault-tolerance mechanisms which partly eliminate the speed-up obtained by this method. Consequently, this approach is only applicable for databases of medium size [28]. An approach presented in [10] uses an index-based near neighbor strategy based on locality sensitive hashing (LSH). Instead of considering long feature sequences, the audio material is split up into small overlapping *shingles* that consist of short chroma feature subsequences. The shingles are then indexed using locality sensitive hashing which allows for scaling this approach to larger datasets. However, to cope with temporal variations, each shingle covers only a small portion of the audio material and queries need to consist of a large number of shingles. The high number of table look-ups induced by this strategy may become problematic for very large datasets where the index is stored on a secondary storage device. The approach presented in [20] is also based on LSH. However, to reduce the number of table look-ups, each query consists of only a single shingle covering 15–25 seconds of the audio. To handle temporal variations, a combination of local feature smoothing and global query scaling is proposed.

In summary, mid-specific audio matching using a combination of highly robust chroma features and sequence-based similarity measures that account for different tempi results in a good retrieval quality. However, the low specificity of this task makes indexing much harder than in the case of audio identification. This task becomes even more challenging when dealing with relatively short fragments on the query and database side.

4 Version Identification

In the previous tasks, a musical fragment is used as query and similar fragments or documents are retrieved according to a given degree of specificity. The degree of specificity was very high for audio identification and more relaxed for audio matching. If we allow for even less specificity, we are facing the problem of version identification [42]. In this scenario, a user wants to retrieve not only exact or near-duplicates of a given query, but also any existing re-interpretation of it, no matter how radical such a re-interpretation might be. In general, a version may differ from the original recording in many ways, possibly including significant changes in timbre, instrumentation, tempo, main tonality, harmony, melody, and lyrics. For example, in addition to the aforementioned Karajan's rendition of Beethoven's Symphony No. 5, one could be also interested in a live performance of it, played by a punk-metal band who changes the tempo in a non-uniform way, transposes the piece to another key, and skips many notes as well as most parts of the original structure. These types of documents where, despite numerous and important variations, one can still unequivocally glimpse the original composition are the ones that motivate version identification.

Version identification is usually interpreted as a document-level retrieval task, where a single similarity measure is considered to globally compare entire documents [3, 13, 46]. However, successful methods perform this global comparison on a local basis. Here, the final similarity measure is inferred from locally comparing only parts of the documents—a strategy that allows for dealing with non-trivial structural changes. This way, comparisons are performed either on some representative part of the piece [18], on short, randomly chosen subsequences of it [32], or on the best possible longest matching subsequence [43, 44].

A common approach to version identification starts from the previously introduced chroma features; also more general representations of the tonal content such as chords or tonal templates have been used [42]. Furthermore, melody-based approaches have been

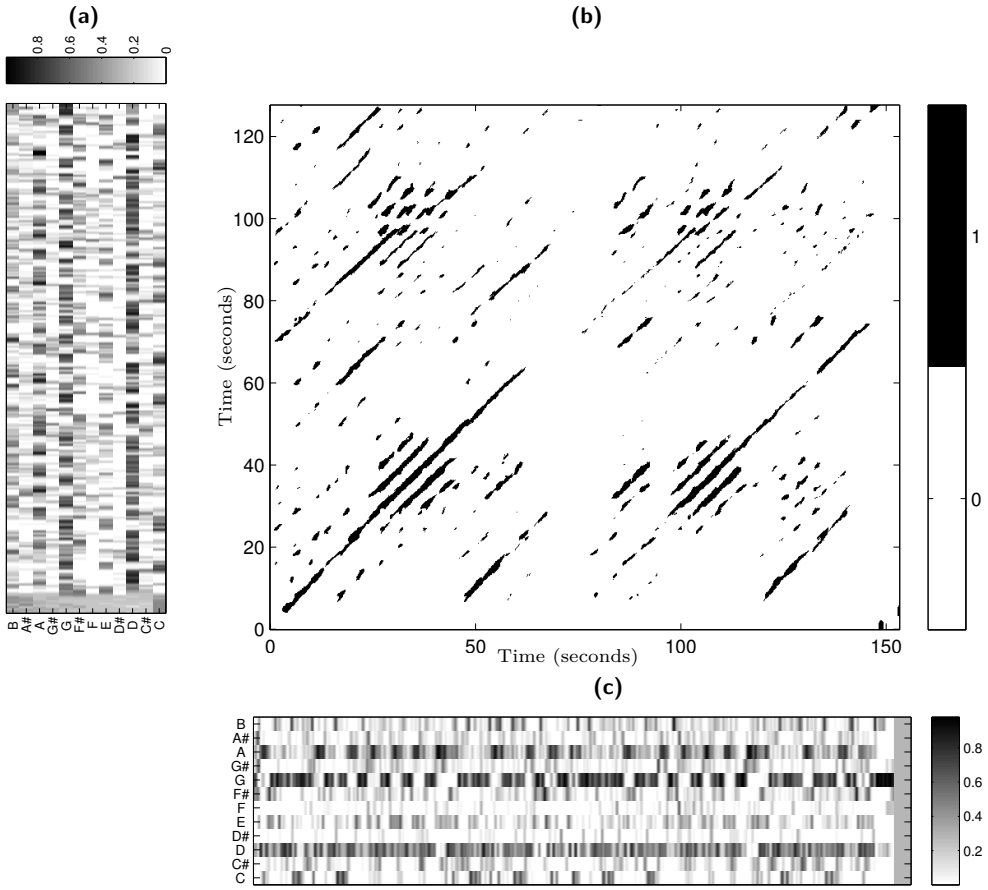


Figure 8 Similarity matrix for "Act naturally" by The Beatles, which is actually a cover version of a song by Buck Owens. **(a)** Chroma features of the version by The Beatles. **(b)** Score matrix. **(c)** Chroma features of the version by Buck Owens.

suggested, although recent findings suggest that this representation may be suboptimal [15, 40]. Once a tonal representation is extracted from the audio, changes in the main tonality need to be tackled, either in the extraction phase itself, or when performing pairwise comparisons of such representations.

Tempo and timing deviations have a strong effect in the chroma feature sequences, hence making their direct pairwise comparison problematic. An intuitive way to deal with global tempo variations is to use beat-synchronous chroma representations [6, 13]. However, the required beat tracking step is often error-prone for certain types of music and therefore may negatively affect the final retrieval result. Again, as for the audio matching task, dynamic programming algorithms are a standard choice for dealing with tempo variations [34], this time applied in a local fashion to identify longest matching subsequences or local alignments [43, 44].

An example of such an alignment procedure is depicted in Figure 8 for our "Act naturally" example by The Beatles. The chroma features of this version are shown in Figure 8c. Actually, this song is originally not written by The Beatles but a cover version of a Buck Owens song of the same name. The chroma features of the original version are shown in Figure 8a. Alignment algorithms rely on some sort of scores (and penalties) for matching (mismatching) individual chroma sequence elements. Such scores can be real-valued or binary. Figure 8b shows a binary score matrix encoding pair-wise similarities between chroma vectors of the two sequences. The binarization of score values provides some additional robustness against

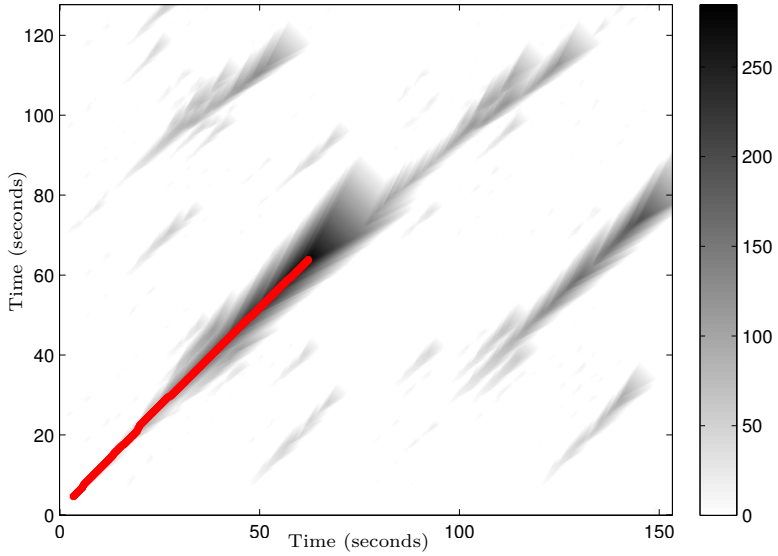


Figure 9 Accumulated score matrix with optimal alignment path for the "Act naturally" example (as shown in Figure 8).

small spectral and tonal differences. Correspondences between versions are revealed by the score matrix in the form of diagonal paths of high score. For example, in Figure 8, one observes a diagonal path indicating that the first 60 seconds of the two versions exhibit a high similarity.

For detecting such path structures, dynamic programming strategies make use of an accumulated score matrix. In their local alignment version, where one is searching for subsequence correspondences, this matrix reflects the lengths and quality of such matching subsequences. Each element (consisting of a pair of indices) of the accumulated score matrix corresponds to the end of a subsequence and its value encodes the score accumulated over all elements of the subsequence. Figure 9 shows an example of the accumulated score matrix obtained for the score matrix in Figure 8. The highest-valued element of the accumulated score matrix corresponds to the end of the most similar matching subsequence. Typically, this value is chosen as the final score for the document-level comparison of the two pieces. Furthermore, the specific alignment path can be easily obtained by backtracking from this highest element [34]. The alignment path is indicated by the red line in Figure 9. Additional penalties account for the importance of insertions/deletions in the subsequences. In fact, the way of deriving these scores and penalties is usually an important part of the version identification algorithms and different variants have been proposed [3, 43, 44]. The aforementioned final score is directly used for ranking candidate documents to a given query. It has recently been shown that such rankings can be improved by combining different scores obtained by different methods [39], and by exploiting the fact that alternative renditions of the same piece naturally cluster together [30, 45].

The task of version identification allows for these and many other new avenues for research [42]. However, one of the most challenging problems that remains to be solved is to achieve high accuracy and scalability at the same time, allowing low-specific retrieval in large music collections [6]. Unfortunately, the accuracies achieved with today's non-scalable approaches have not yet been reached by the scalable ones, the latter remaining far behind the former.

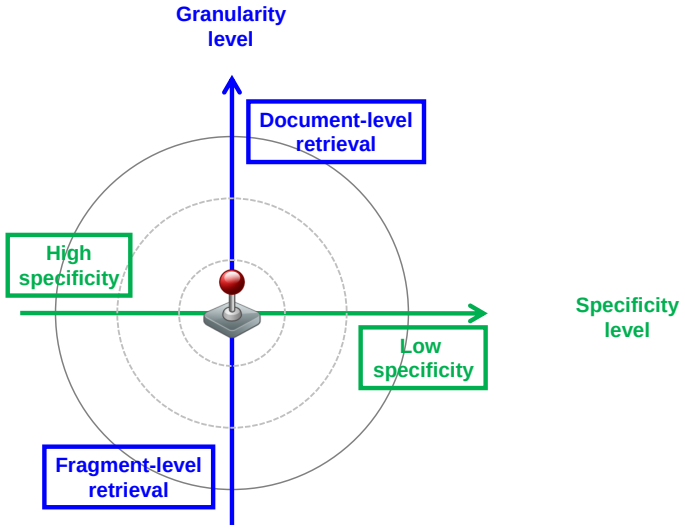


Figure 10 Joystick-like user interface for continuously adjusting the specificity and granularity levels used in the retrieval process.

5 Outlook

In this paper, we have discussed three representative retrieval strategies based on the query-by-example paradigm. Such content-based approaches provide mechanisms for discovering and accessing music even in cases where the user does not explicitly know what he or she is actually looking for. Furthermore, such approaches complement traditional approaches that are based on metadata and tags. The considered level of specificity has a significant impact on the implementation and efficiency of the retrieval system. In particular, search tasks of high specificity typically lead to exact matching problems, which can be realized efficiently using indexing techniques. In contrast, search tasks of low specificity need more flexible and cost-intensive mechanisms for dealing with spectral, temporal, and structural variations. As a consequence, the scalability to huge music collections comprising millions of songs still poses many yet unsolved problems.

Besides efficiency issues, one also has to better account for user requirements in content-based retrieval systems. For example, one may think of a comprehensive framework that allows a user to adjust the specificity level at any stage of the search process. Here, the system should be able to seamlessly change the retrieval paradigm from high-specific audio identification, over mid-specific audio matching and version identification to low-specific genre identification. Similarly, the user should be able to flexibly adapt the granularity level to be considered in the search. Furthermore, the retrieval framework should comprise control mechanisms for adjusting the musical properties of the employed similarity measure to facilitate searches according to rhythm, melody, or harmony or any combination of these aspects.

Figure 10 illustrates a possible user interface for such an integrated content-based retrieval framework, where a joystick allows a user to continuously and instantly adjust the retrieval specificity and granularity. For example, a user may listen to a recording of Beethoven's Symphony No. 5, which is first identified to be a Bernstein recording using an audio identification strategy (moving the joystick to the leftmost position). Then, being interested in different

versions of this piece, the user moves the joystick upwards (document-level) and to the right (mid-specific), which triggers a version identification. Subsequently, shifting towards a more detailed analysis of the piece, the user selects the famous fate motif as query and moves the joystick downwards to perform some mid-specific fragment-based audio matching. Then, the system returns the positions of all occurrences of the motif in all available interpretations. Finally, moving the joystick to the rightmost position, the user may discover recordings of pieces that exhibit some general similarity like style or mood. In combination with immediate visualization, navigation, and feedback mechanisms, the user is able to successively refine and adjust the query formulation as well as the retrieval strategy, thus leading to novel strategies for exploring, browsing, and interacting with large collections of audio content.

Another major challenge refers to cross-modal music retrieval scenarios, where the query as well as the retrieved documents can be of different modalities. For example, one might use a small fragment of a musical score to query an audio database for recordings that are related to this fragment. Or a short audio fragment might be used to query a database containing MIDI files. In the future, comprehensive retrieval frameworks are to be developed that offer multi-faceted search functionalities in heterogeneous and distributed music collections containing all sorts of music-related documents.

6 Acknowledgment

We would like to express our gratitude to Christian Dittmar, Emilia Gómez, Frank Kurth, and Markus Schedl for their helpful and constructive feedback.

References

1 Eric Allamanche, Jürgen Herre, Oliver Hellmuth, Bernhard Fröba, and Markus Cremer. AudioID: Towards content-based identification of audio material. In *Proceedings of the 110th AES Convention*, Amsterdam, NL, 2001.

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

3 Juan Pablo Bello. Audio-based cover song retrieval using approximate chord sequences: testing shifts, gaps, swaps and beats. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 239–244, Vienna, Austria, 2007.

4 Thierry Bertin-Mahieux, Douglas Eck, Francois Maillet, and Paul Lamere. Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, 37(2):115–135, 2008.

5 Thierry Bertin-Mahieux, Douglas Eck, and Michael I. Mandel. Automatic tagging of audio: The state-of-the-art. In Wenwu Wang, editor, *Machine Audition: Principles, Algorithms and Systems*, chapter 14, pages 334–352. IGI Publishing, 2010.

6 Thierry Bertin-Mahieux and Daniel P.W. Ellis. Large-scale cover song recognition using hashed chroma landmarks. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Platz, NY, 2011.

7 Dmitry Bogdanov, Joan Serrà, Nicolas Wack, and Perfecto Herrera. Unifying low-level and high-level music similarity measures. *IEEE Transactions on Multimedia*, 13(4):687 –701, aug. 2011.

8 Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma. A review of audio fingerprinting. *Journal of VLSI Signal Processing Systems*, 41(3):271–284, 2005.

9 Pedro Cano, Eloi Batlle, Harald Mayer, and Helmut Neuschmied. Robust sound modeling for song detection in broadcast audio. In *Proceedings of the 112th AES Convention*, pages 1–7, 2002.

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

11 Michael A. Casey, Remco Veltkap, 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, 2008.

12 Òscar Celma. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer, 1st edition, September 2010.

13 Daniel P.W. Ellis and Graham E. Poliner. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 1429–1432, Honolulu, Hawaii, USA, April 2007.

14 Sébastien Fenet, Gaël Richard, and Yves Grenier. A scalable audio fingerprint method with robustness to pitch-shifting. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, Miami, USA, 2011.

15 Rémi Foucard, Jean-Louis Durrieu, Mathieu Lagrange, and Gaël Richard. Multimodal similarity between musical streams for cover version detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5514–5517, Dallas, USA, 2010.

16 Dimitrios Fragoulis, George Rousopoulos, Thanasis Panagopoulos, Constantin Alexiou, and Constantin Papaodysseus. On the automated recognition of seriously distorted musical recordings. *IEEE Transactions on Signal Processing*, 49(4):898–908, 2001.

17 Emilia Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, UPF Barcelona, 2006.

18 Emilia Gómez, Bee Suan Ong, and Perfecto Herrera. Automatic tonal analysis from music summaries for version identification. In *Proceedings of the 121st AES Convention*, San Francisco, CA, USA, 2006.

19 Fabien Gouyon. *A computational approach to rhythm description: audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2005.

20 Peter Grosche and Meinard Müller. Toward characteristic audio shingles for efficient cross-version music retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 473–476, Kyoto, Japan, 2012.

21 Peter Grosche and Meinard Müller. Toward musically-motivated audio fingerprints. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 93–96, Kyoto, Japan, 2012.

22 Jaap Haitsma and Ton Kalker. A highly robust audio fingerprinting system. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 107–115, Paris, France, 2002.

23 Jaap Haitsma and Ton Kalker. Speed-change resistant audio fingerprinting using auto-correlation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 728–731, 2003.

24 Yan Ke, Derek Hoiem, and Rahul Sukthankar. Computer vision for music identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 597–604, San Diego, CA, USA, 2005.

25 Hyoung-Gook Kim, Nicolas Moreau, and Thomas Sikora. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. John Wiley & Sons, 2005.

26 Youngmoo E. Kim, Erik M. Schmidt, Raymond Migneco, Brandon C. Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A. Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 255–266, Utrecht, The Netherlands, 2010.

27 Aniket Kittur, Ed Chi, Bryan A. Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Computer/Human Interaction Conference (Alt.CHI)*, San Jose, CA, 2007.

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

29 Frank Kurth, Andreas Ribbrock, and Michael Clausen. Identification of highly distorted audio material for querying large scale data bases. In *Proceedings of the 112th AES Convention*, 2002.

30 Mathieu Lagrange and Joan Serrà. Unsupervised accuracy improvement for cover song detection using spectral connectivity network. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 595–600, 2010.

31 Paul Lamere. Social tagging and music information retrieval. *Journal of New Music Research*, 37(2):101–114, 2008.

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

33 Matthias Mauch and Simon Dixon. Approximate note transcription for the improved identification of difficult chords. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 135–140, 2010.

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

35 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.

36 Meinard Müller and Sebastian Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 215–220, Miami, USA, 2011.

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

38 Mathieu Ramona and Geoffroy Peeters. Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 477–480, 2011.

39 Suman Ravuri and Daniel P.W. Ellis. Cover song detection: from high scores to general classification. In *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, pages 65–68, Dallas, TX, 2010.

40 Justin Salamon, Joan Serrà, and Emilia Gómez. Melody, bass line and harmony descriptions for music version identification. In *preparation*, 2011.

41 Björn Schuller, Florian Eyben, and Gerhard Rigoll. Tango or waltz?: Putting ballroom dance style into tempo detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2008:12, 2008.

42 Joan Serrà, Emilia Gómez, and Perfecto Herrera. Audio cover song identification and similarity: background, approaches, evaluation and beyond. In Z. W. Ras and A. A. Wieczorkowska, editors, *Advances in Music Information Retrieval*, volume 16 of *Studies in Computational Intelligence*, chapter 14, pages 307–332. Springer, Berlin, Germany, 2010.

43 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, oct 2008.

44 Joan Serrà, Xavier Serra, and Ralph G. Andrzejak. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9):093017, 2009.

45 Joan Serrà, Massimiliano Zanin, Perfecto Herrera, and Xavier Serra. Characterization and exploitation of community structure in cover song networks. *Pattern Recognition Letters*, 2010. Submitted.

46 Wei-Ho Tsai, Hung-Ming Yu, and Hsin-Min Wang. Using the similarity of main melodies to identify cover versions of popular songs for music document retrieval. *Journal of Information Science and Engineering*, 24(6):1669–1687, 2008.

47 Emiru Tsunoo, Taichi Akase, Nobutaka Ono, and Shigeki Sagayama. Musical mood classification by rhythm and bass-line unit pattern analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010.

48 Douglas Turnbull, Luke Barrington, and Gert Lanckriet. Five approaches to collecting tags for music. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 225–230, Philadelphia, USA, 2008.

49 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, 2008.

50 George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

51 Jan Van Balen. Automatic recognition of samples in musical audio. Master's thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2011.

52 Avery Wang. An industrial strength audio search algorithm. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 7–13, Baltimore, USA, 2003.

53 Felix Weninger, Martin Wöllmer, and Björn Schuller. Automatic assessment of singer traits in popular music: Gender, age, height and race. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 37–42, Miami, Florida, USA, 2011.

54 Kris West and Paul Lamere. A model-based approach to constructing music similarity functions. *EURASIP Journal on Advances in Signal Processing*, 2007(1):024602, 2007.

Data-Driven Sound Track Generation*

Meinard Müller and Jonathan Driedger

Saarland University and MPI Informatik
Campus E1-4, 66123 Saarbrücken, Germany
meinard@mpi-inf.mpg.de, driedger@mpi-inf.mpg.de

Abstract

Background music is often used to generate a specific atmosphere or to draw our attention to specific events. For example in movies or computer games it is often the accompanying music that conveys the emotional state of a scene and plays an important role for immersing the viewer or player into the virtual environment. In view of home-made videos, slide shows, and other consumer-generated visual media streams, there is a need for computer-assisted tools that allow users to generate aesthetically appealing music tracks in an easy and intuitive way. In this contribution, we consider a data-driven scenario where the musical raw material is given in form of a database containing a variety of audio recordings. Then, for a given visual media stream, the task consists in identifying, manipulating, overlaying, concatenating, and blending suitable music clips to generate a music stream that satisfies certain constraints imposed by the visual data stream and by user specifications. It is our main goal to give an overview of various content-based music processing and retrieval techniques that become important in data-driven sound track generation. In particular, we sketch a general pipeline that highlights how the various techniques act together and come into play when generating musically plausible transitions between subsequent music clips.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems

Keywords and phrases Sound track, content-based retrieval, audio matching, time-scale modification, warping, tempo, beat tracking, harmony

Digital Object Identifier 10.4230/DFU.Vol3.11041.175

1 Introduction

The computer-assisted generation of sound tracks for given visual media streams has significantly gained in importance. For example, video games of these days are often accompanied by music of high artistic value and excellent sound quality coming close to sound tracks of movies. However, opposed to film music, the sound track underlying a video game has to constantly adapt to the respective scene of the game and to interactively react to the player's input.

When developing a high-quality computer game, composers are asked to create specific music clips that not only match the various scenes and characters of the game, but also account for transitions within and across different scenes. To this end, the music needs to contain various transition points that allow for smoothly connecting and bridging different passages at specified or even arbitrary points in time. Even though there may be no real

* This work has been supported by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) and the German Research Foundation (DFG MU 2682/5-1). Meinard Müller is now with Bonn University, Department of Computer Science III, Germany.

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 175–194

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

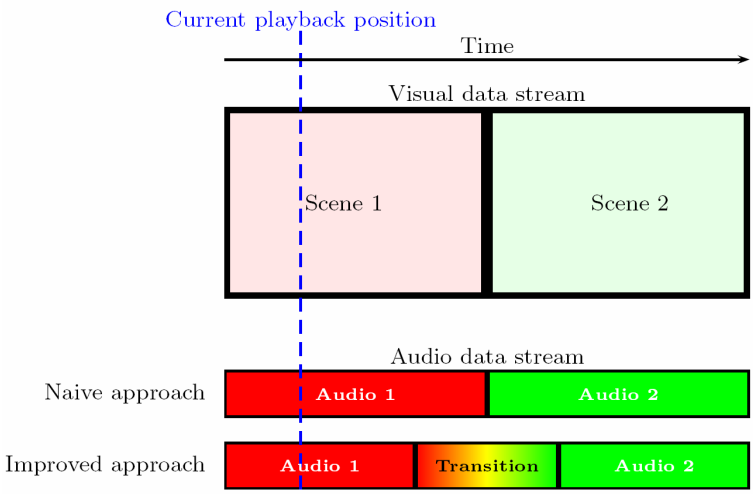


Figure 1 Sound track generation by concatenating existing audio clips.

alternatives to manually creating music in particular when artistic aspects are given priority, such *compositional approaches* to sound track generation are costly and labor extensive. Furthermore, the resulting music is highly specialized, and the system has a slow response time when transitions are possible only at specific pre-defined positions.

As an inexpensive alternative, one may revert to *parametric approaches*, where the background music is synthesized based on parametric models. Here, the free parameters allow for specifying, adjusting, and triggering sound events and may directly be controlled by scene annotations, by the moving objects within the scene, or by a user's input. However, even though offering fast response times, such parametric approaches may be aesthetically questionable from a musical point of view.

The main focus of this work follows a third approach. To obtain appealing sound tracks, one strategy is to simply play back an existing music recording that is in line with a given visual data stream. Reverting to an audio database comprising high-quality music recordings, the idea of such a *data-driven approach* is to identify and play back music clips that correspond well to the visual scenes while accounting for user specifications. However, a simple concatenation of audio clips may result in unpleasant and abrupt transitions between subsequent audio clips. Therefore, one main challenge consists in the creation of musically smooth and euphonious transitions, which are as pleasant as possible to the listener's ear, see Figure 1.

In this contribution, our main goal is to describe a possible pipeline for such a data-driven sound track generation system while giving an overview of the necessary data processing and retrieval techniques, see Figure 2. In the following, we exemplarily consider an online scenario, where a visual data stream, which consists of a sequence of changing scenes that are associated to certain categories (e. g., moods), is given. Furthermore, a comprehensive music database that contains audio recordings of various genres, styles and moods serves as basis for the sound track to be generated. These recordings are assumed to be annotated with respect to the same categories as used to describe the visual scenes. For the current scene, a specific audio recording is played back. As soon as the next scene change is pending, the category of the subsequent scene as well as the tolerable delay for the transition needs to be known. The system then determines a suitable region in the current audio recording,

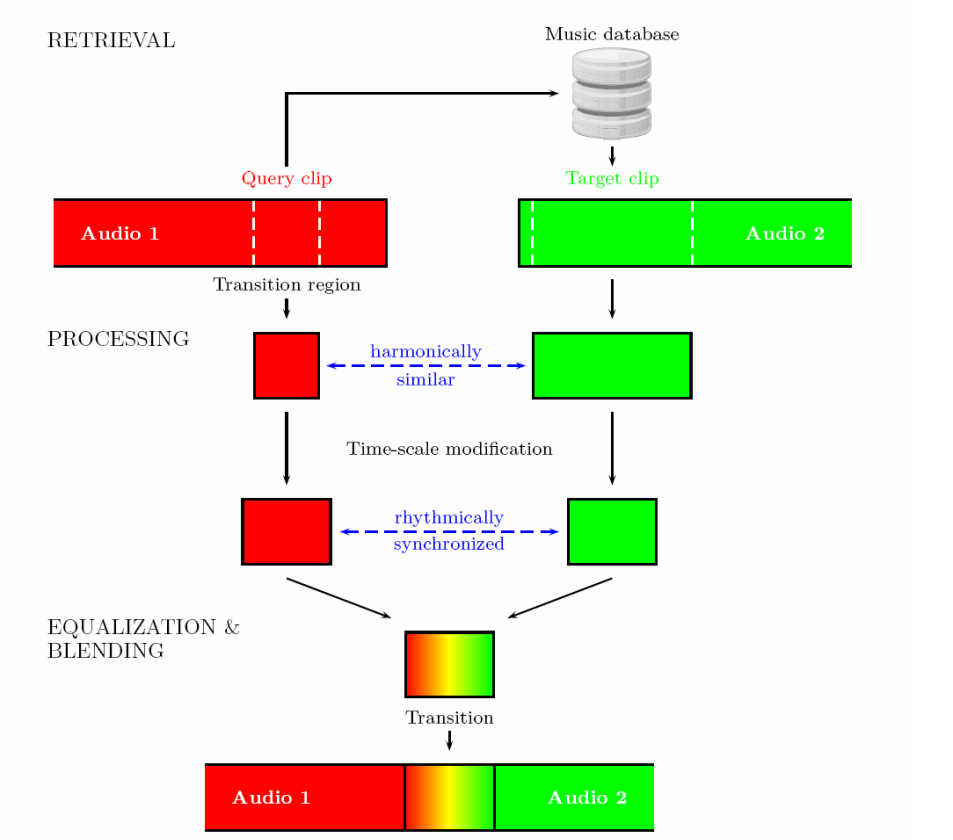


Figure 2 Overview of different retrieval and processing components required for a data-driven sound track generation system.

also referred to as *transition region*. The waveform corresponding to this region is then used as *query clip*, and content-based retrieval is performed to identify a suitable audio clip in the music database—referred to as *target clip*—satisfying the following two properties. Firstly, the clip should be contained in an audio recording that reflects the category of the subsequent scene. Secondly, the target clip should be similar to the query clip to allow for a smooth (e. g., harmonically and rhythmically plausible) transition. To this end, one particularly needs a clip that has a similar harmonic progression as the query clip. In the next step, the two clips are temporally synchronized by first estimating the beat positions and then applying suitable time-scale modifications (similar to what a DJ is doing). The actual transition from the current recording (containing the query clip) to the next recording (containing the target clip) is then realized by blending from the synchronized query clip to the target clip. Finally, to further improve the quality of the transition, one needs intelligent equalization techniques that can be used to attenuate possibly interfering sound components or to amplifying certain voices, instruments or notes.

In the sketched approach, various challenges arise. First, one needs similarity measures and content-based retrieval strategies to search for and identify suitable music clips that satisfy the given constraints. These constraints may not only be imposed by the visual input and user specifications, but also by algorithmic and aesthetic considerations. Furthermore, one requires a number of signal processing techniques that allow for adjusting the audio

material with respect to various musical aspects including harmony, rhythm, tempo, or polyphony. In the following, we give an overview of these techniques and provide suitable links to the literature. The remainder of this contribution is organized as follows. In Section 2, we discuss previous work that is related to the problem of automated sound track generation. Then, in Section 3, we give an overview of the involved data processing and retrieval techniques while highlighting how these techniques act together and come into play in a data-driven soundtrack generation scenario. Finally, we conclude with Section 4 discussing challenges, limitations, and future work.

2 Related Work

The idea of generating new music by concatenating existing music fragments based on euphonious transitions has a long history. At the end of the 18$^{\text{th}}$ century, "Musikalische Würfelspiele" ("Musical dice games") were a popular pastime, where a piano player had to create music by suitably concatenating measures from known pieces that were randomly chosen by throwing a dice [32, 22].

Nowadays the generation of dynamically changing music by concatenating pre-rendered music clips has become an important issue in particular in the context of video games. As emphasized in [70], the generation of suitable music can add emotional depth and soul to the various scenes leading to a highly immersive gaming experience. To this end, the music not only has to loosely reflect the mood of the respective scene, but also has to constantly adapt to, or even to anticipate the game's events and the player's actions. The term *adaptive audio* (or *adaptive music*) has been used to describe audio and music that responds appropriately to gameplay [70]. As one requirement, to support the game's continuity, music transitions that seamlessly connect the various moods and intensities are needed. To this end, one needs techniques that go far beyond a simple concatenation or cross-fade between subsequent audio clips. Instead, short building blocks of music, different layers (e. g., percussion loops, super-imposable melody and instrument tracks), as well as transitional cues are required for creating adaptive music. The composition of music that does not follow a linear flow (as is for traditional music) but that can be reassembled in a flexible and smooth fashion constitutes a hard problem—musically as well as technically. For a detailed discussion and further links, we also refer to [66].

There are various approaches to automatically generate music streams on the basis of *symbolic* music representations. For example, [11] describes an automated music generation system that works on the basis of MIDI files. Opposed to waveform-based audio representations, symbolic representations offer more flexibility and direct control since musical parameters such as note events, instrumentation, or tempo are given explicitly and can be therefore altered easily. On the downside, synthesizing music from a symbolic representation often leads to unsatisfying results, e. g., because of the artificiality of the used synthetic instruments or the lacking of performance nuances. Furthermore, high-quality symbolic representations are often not available or hard to generate from existing audio material.

The automated remixing and concatenation of existing audio material constitutes a challenging area of research. A prominent application scenario is what a disc jockey (DJ) is typically doing: he not only selects appropriate music for the audience, but also tries to mix and blend recorded music to create a continuous playback. First systems to automate this process are described, e. g., in [8, 37]. In the mixing process, DJs pay particular attention to a good rhythmical transition, which requires an adjustment of the tempo and a matching of the beats. A tool to automate the process of finding good rhythmical transitions is described

in [39]. Harmonic similarity of the two audio clips to be connected usually plays a minor role, even though professional DJs also often try to take the musical key into consideration. In [49], the authors describe a first system for concatenating audio clips to form a single long audio stream, where the recordings are ordered in such a way that euphonious transitions between the clips are possible. The positions of the transitions are chosen to maximize the local harmonic and rhythmic similarity of the two subsequent audio clips. This scenario is similar to what we want to consider in our contribution. However, we want to focus more on the underlying techniques that are need for realizing such a framework, whereas [49] describe a first overall system. Finally, we want to mention the work by [69], where the goal is to temporally rearrange a given music recording to fit certain user-specified constraints. In particular, suitable transitions points are identified within the audio material that allow for deleting, copying, and rearranging certain parts while keeping the flow of the music. This not only allows for an automated adjustment of the duration of a given recording but also for linking certain parts of the recording to specified key frames of the visual data stream.

3 Music Retrieval and Processing

We now give an overview of the various content-based music retrieval and processing techniques that are important in view of the described data-driven sound track generation scenario, see also Figure 12 for an overview. In particular, we have a focus on the creation of musically plausible transitions between audio recordings that are to be concatenated. To this end, one requires methods from audio signal analysis to capture harmonic and rhythmic properties of music recordings. Such properties form the basis for designing musically meaningful similarity measures needed to identify potential transition regions. Then, one requires manipulation techniques that allow for temporally (e. g., time-scaling, clipping) and spectrally (e. g., modulation, harmonic-percussive separation, voice equalization) manipulating the audio material. Furthermore, synthesis methods (e. g., blending, morphing) are needed to render the final audio stream. Last but not least, in view of efficiency and online capability, data structures are to be developed that facilitate fast content-based search and encode, for example, plausible transitions between music clips.

3.1 Category-based Classification

As mentioned in the introduction, we assume that the visual scenes are associated to certain categories that may refer to the emotional content or mood of the scene. For example, the current scene may be associated with the attribute "Angry" whereas the subsequent scene may by associated with the attribute "Happy." Then one important step in the sound track generation scenario is to find music recordings that reflect the categories of the given scenes, see Figure 3.

Actually, the automated classification of music recordings with respect to a given set of categories has been a central topic in the field of music information retrieval. Generally, such categories refer to cultural or musicological aspects [16] including genre [59, 63] or rhythm styles [26, 61]. In our scenario, we are particularly interested in categories that refer to mood or emotions [36, 41, 62]. However, as noted in [41], when organizing music in terms of emotional content, one is faced with the problem that there is a "considerable disagreement regarding the perception and interpretation of the emotions of a song or ambiguity within the piece itself." In other words, the categories are often ill-defined and highly subjective with the result that the automation of the classification problem is still in its early stages, see [41] for an overview.

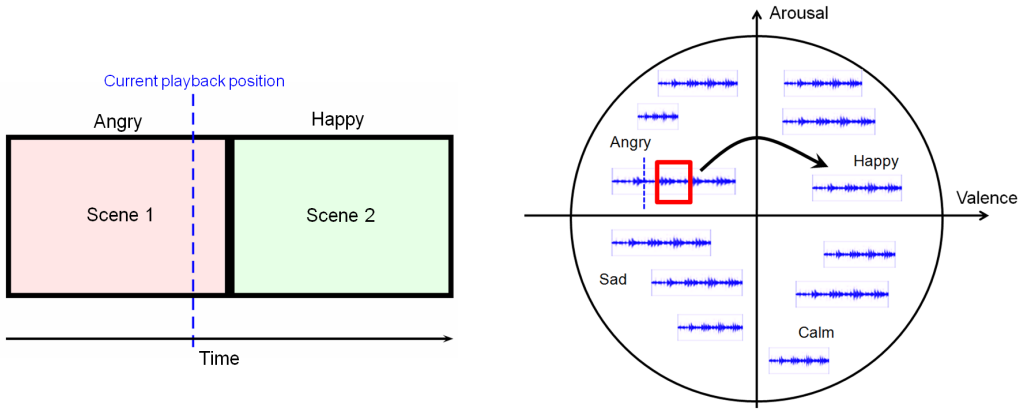


Figure 3 Visual scenes and music database annotated with respect to mood categories of the valence-arousal space [58].

In the following, we assume that the recordings of the database have been annotated according to given mood categories. Such annotations may be obtained by manual expert classification or may be derived from contextual text information (e. g., websites, tags, and lyrics) and content-based approaches [41]. In the music context, the most prominent way to organize emotional descriptors is the two-dimensional valence-arousal space as originally introduced in [58], see Figure 3. Here, the mood categories are arranged on a plane with two independent axes that encode arousal (intensity) ranging from low to high and valence (appraisal of polarity) ranging from negative to positive [41]. However, the specific nature of the descriptive labels and their organization is not in the scope of this contribution. In the following, we only require that both the visual scenes as well as the database documents are characterized based on the same set of categories.

3.2 Content-based Audio Retrieval

In our online scenario, we assume that a music recording is played back underlying the current visual scene. Once a scene change is pending and the category of the subsequent scene is known, the goal is to find a music recording which category fits the subsequent scene. Assuming suitable annotations as discussed in Section 3.1, this simply requires a table look-up to retrieve all music documents of the desired category. In addition, we want to generate a smooth transition from the current recording to the next one. Here, a simple cross-fade between two recordings may result in unpleasant listening experiences due to harmonic, melodic and rhythmic incompatibilities in the transition phase. Instead, the goal is to generate musically transitions that do not intercept the flow of the multimedia presentation. One way to achieve this goal is to specify a suitable region in the current audio recording, where the transition to the next recording is to be performed. Based on the corresponding clip, one then needs to identify a recording that contains a semantically related target clip allowing for a plausible transition. This is exactly the point, where content-based audio retrieval comes into play. In the following, we summarize two prominent retrieval scenarios and describe the techniques used in our pipeline.

Actually, in content-based audio retrieval, various levels of specificity can be considered. At the highest specificity level, the retrieval task is often referred to as *audio identification* or *audio fingerprinting*. Here, given a small audio fragment as query, the task consists in

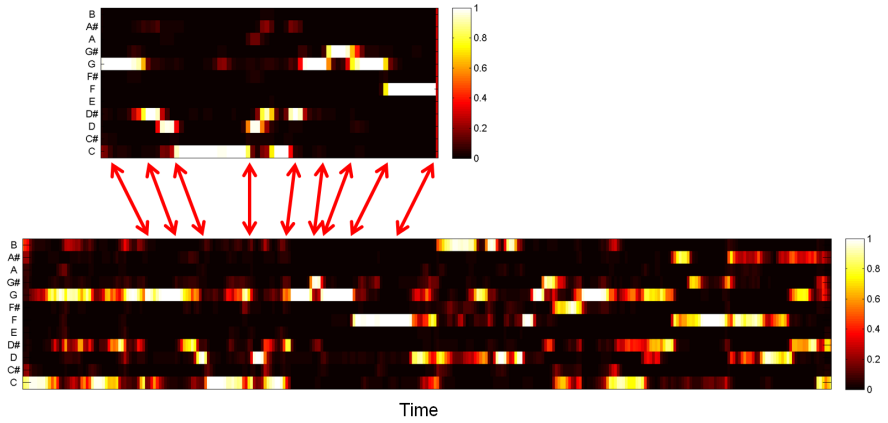


Figure 4 Chroma-based audio matching procedure. The red arrows indicate temporal correspondences between the query clip and a local section of a given music recording.

identifying the fragment (i. e., retrieving the audio recording containing the fragment along with the fragment's position) within a large audio collection [1, 5, 46, 68]. Note that at this level, the notion of similarity is rather close to the identity. Even though recent identification algorithms show a significant degree of robustness towards noise, MP3 compression artifacts, and uniform temporal distortions, existing algorithms for audio identification can not deal with strong non-linear temporal distortions or with other musically motivated variations that concern, for example, the articulation or instrumentation.

In sound track generation scenarios as described in [69], where the goal is to identify possible transition points within the same music recording, such strict notions of similarity may be meaningful. However, when changing from one music recording to a completely different one, a much coarser notion of similarity to identify potential transition regions is required. The identification of such regions can be accomplished by using *audio matching* techniques, where the goal is to retrieve all audio clips that musically correspond to the query [54, 50, 45]. In audio matching, opposed to traditional audio identification, one allows variations in musical aspects such as tempo, instrumentation, loudness, timbre, or accentuation.

In our proposed pipeline, we are specifically looking for audio clips that are harmonically related. Therefore, we use a chroma-based audio matching procedure as originally described in [54]. The general idea is to convert the audio material into mid-level representations that show a high degree of robustness to variations that are to be left unconsidered in the comparison. On the other hand, the feature representations should capture characteristic information that musically relate the identified clips to facilitate a plausible transition. In this context, *chroma-based audio features* have turned out to be a suitable mid-level representation [3, 24, 51]. Assuming the equal-tempered scale, the chroma attributes correspond to the set $\{\mathrm{C}, \mathrm{C}^\sharp, \mathrm{D}, \ldots, \mathrm{B}\}$ that consists of the twelve pitch spelling attributes as used in Western music notation. Representing the short-time content of a music representation in each of the 12 pitch classes, chroma features[1] show a large degree of robustness to variations in timbre and dynamics, while keeping sufficient information to characterize the rough harmonic

[1] MATLAB implementations for some chroma variants are supplied by the Chroma Toolbox: `www.mpi-inf.mpg.de/resources/MIR/chromatoolbox`, see also [53]

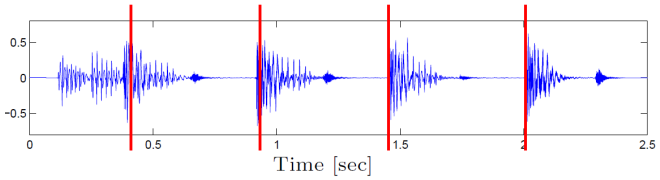


Figure 5 Beat tracking result (indicated by the red vertical lines) for a given music recording.

progression of the underlying piece of music. Based on these feature representations, the query clip is locally compared with clips that are contained in the target music recordings using alignment techniques. In particular, we use a local variant of *Dynamic Time Warping* (DTW) that can be used to find optimal temporal correspondences between the query clip and a local section of a given music recording [51]. Intuitively, these correspondences can be thought of a linking structure as indicated by the red arrows shown in Figure 4. These arrows encode how the feature sequences are to be warped (in a non-linear fashion) to match each other.

In [54], the main application of audio matching is to identify different versions of the same piece of music irrespective of the performance, instrumentation, or arrangement. As reported in [45, 29], using a query length of roughly 20 seconds (or more) leads to a high precision for this task. Now, in the sound track generation scenario as tackled in this paper, one is typically not interested in different versions of the same piece of music, but in harmonically related passages contained in *different* pieces. Such passages can be obtained when using query clips of shorter duration (less than 10 seconds). In other words, what is considered a false positive match in [54], may be a desirable match in our scenario.

3.3 Tempo and Beat Tracking

The chroma-based audio matching procedure is used to identify a target audio clip that shares a similar harmonic progression with the query clip. In view of a rhythmically plausible transition, one also needs to temporally synchronize the two clips—similar to what a DJ is doing when matching the beats of two recordings. This leads us to further central tasks referred to as *tempo estimation* and *beat tracking*, where the objective is to automatically locate the beat positions within a given music recording, see Figure 5.

Most approaches to tempo estimation and beat tracking proceed in two steps. In the first step, positions of note onsets within the music signal are estimated. Here, most approaches capture changes of the signal's energy or spectrum and derive a so-called *novelty curve*. The peaks of such a curve yield good indicators for note onset candidates [4, 9]. In the second step, the novelty curve is analyzed to detect reoccurring patterns and quasi-periodic pulse trains [12, 17, 28, 57, 60, 72].

Even though most humans are able to tap along the musical beat when listening to a piece of music, transferring this cognitive process into an automated system that reliably works for a large variety of musical styles is a challenging task. In particular, beat tracking becomes hard in the case that a music recording reveals significant tempo changes. This typically occurs in expressive performances of classical music as a result of ritardandi, accelerandi, fermatas, and rubato [30]. Furthermore, the extraction problem is complicated by the fact that the notions of tempo and beat may not be clearly defined due to a complex hierarchical structure of the rhythm [56]. In particular, there are various levels that are presumed to contribute to the human perception of tempo and beat. All these difficulties and ambiguities

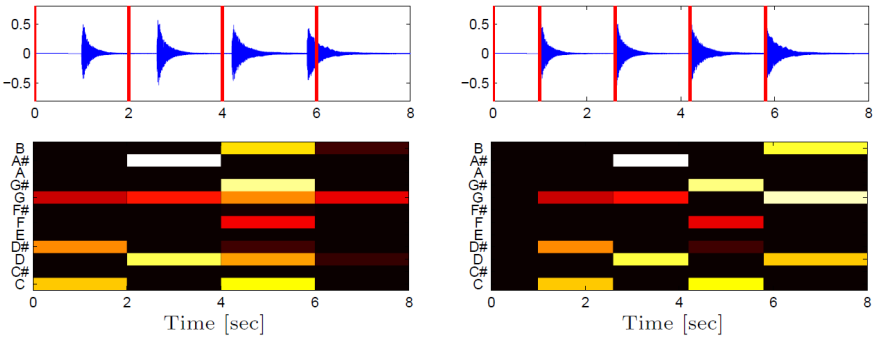


Figure 6 Waveforms and chroma representations using a fixed-size windowing strategy (left) and an adaptive windowing strategy using beat-synchronized windows (right).

have to be kept in mind when using the beat tracking results obtained from automated methods.

Knowing beat positions is not only necessary to temporally synchronize the query and target clip, as will be explained in Section 3.4, but is also beneficial for the feature computation and matching step as we now explain in more detail. When transforming a waveform into some feature representation, one typically splits up the signal into frames using a window function of fixed size and then applies the transform to each frame. Each feature value represents a local property averaged over the respective time window, which may result in "noisy" features when the signal's changes occur within a given window. As an alternative to fixed-size windowing, one can employ a musically more meaningful adaptive windowing strategy, where window boundaries are induced by previously extracted onset and beat positions. Since musical changes typically occur at onset positions, this often leads to an increased homogeneity within the adaptively determined frames which often improves the resulting feature representation, see Figure 6 for an illustration. One major advantage of using *beat-synchronized* audio features is that tempo differences between musically related audio clips are compensated [18]. This alleviates the requirement of using cost-intensive alignment procedures in the retrieval step as discussed in Section 3.2. Furthermore, knowing the beat positions allows for converting a physical time axis (given in seconds) into a musically meaningful time axis (given in beats or measures), which has huge benefits for presenting and comparing music analysis results [43]. However, when relying on beat-synchronous features, one should keep in mind that the quality of automatically extracted beats may be rather poor for certain types of music [30].

3.4 Time-Scale Modification

Once the beat positions are known in the query and target clip, one needs techniques that allow for locally speeding up or slowing down a music recording without changing other characteristics such as the pitch. Originally introduced for speech signals, there are numerous time stretching or *time-scale modification* (TSM) procedures. Most of these procedures are based on a fundamental technique referred to as *Overlap-and-Add* (OLA). The idea is to generate local copies of audio segments, which are obtained by windowing the original audio signal using suitably shifted Hann windows. These copies are then added up (using a constant window overlap) to produce the time-scale modified signal, see Figure 7 for an illustration. Generally, this simple procedure often results in severe noise-like phase distortions and stuttering artifacts which strongly downgrade the quality of the music signal.

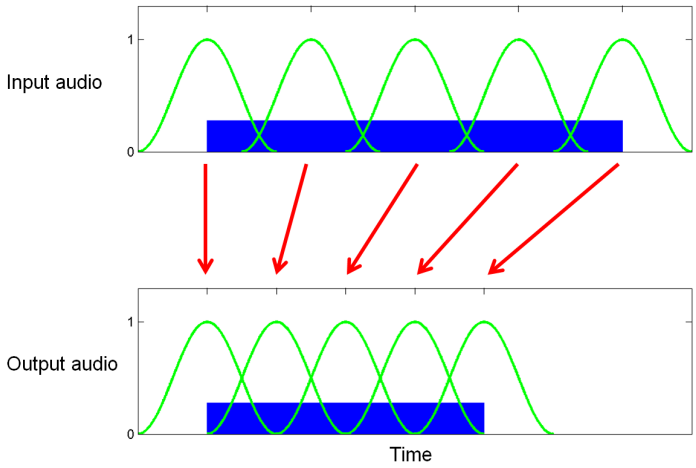


Figure 7 Illustration of the Overlap-and-Add (OLA) technique with blue indicating the waveforms and green indicating the windows.

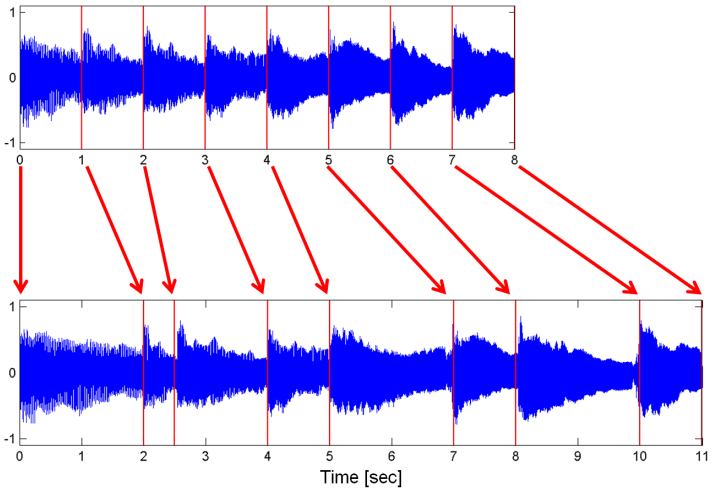


Figure 8 Non-linear time-scale modification of a music recording to temporally adjust beat positions.

Various time-scale modification algorithms have been proposed that try to attenuate these distortions. In general, one can distinguish between time-domain and frequency-domain approaches. A widely used time-domain procedure is known as *WSOLA* (waveform-similarity-based overlap-add) algorithm [65]. Here, phase discontinuities in the fundamental frequency are prevented by slightly adapting the window positions to obtain the local copies using correlation measures before the accumulation step is applied. On the other side, the most common frequency-domain approach is known as *phase vocoder* [13], where one first generates local copies as in the OLA procedure. Next, the phases of each local copy are adjusted in the Fourier domain to achieve a frequency-wise phase coherence in the subsequent accumulation step. To cope with various kinds of artifacts, numerous variants and hybrid methods have been proposed, see, e. g., [2, 14, 15, 25, 27, 40].

In our sound track generation scenario it is of particular importance that the used TSM procedure is capable of performing non-linear time-scale modifications. This is for example

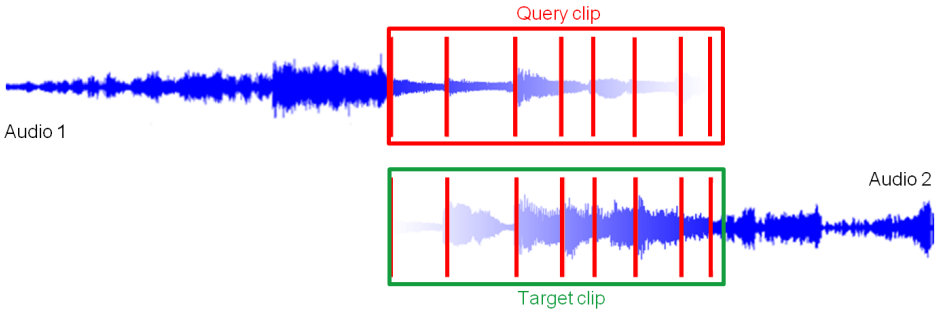


Figure 9 Cross-fade between beat-synchronized query clip and target clip.

needed when adjusting the beat grid of a music recording as shown in Figure 8. Finally, we note that the problem of pitch shifting with the objective to change the pitch of an input signal without changing its duration is dual to the time stretching problem. Here, to shift the pitch of a signal, one can first apply a time-scale modification procedure to stretch the signal and then use a simple sample rate conversion.

3.5 Intelligent Equalization and Blending

After the harmonically related query and target clips have been rhythmically synchronized, one can compute a transition by applying a simple cross-fade between these two clips. Then the transit from the current recording to the subsequent recording can be accomplished smoothly using this transition, see Figure 9.

So far, harmonic and rhythmic aspects were used for retrieving and adjusting the query and target clips. There are many more musical aspects such as instrumentation, musical voices or melodic structures one may want to consider in the transition. To this end, one requires techniques that allow for manipulating the audio material with regard to such aspects. This leads us to another fundamental and challenging area of signal processing generally referred to as *source separation*, where the goal is to decompose a given mixed audio signal into its individual sound sources.

In the musical context, source separation often deals with automatically extracting individual tracks that correspond to different instruments or musical voices from a given audio recording, see [10, 52, 67] for an overview. A related task is to parameterize an audio recording of a piece of music, where the parameters encode musical aspects such as pitch, onset positions, note durations, as well as tuning and timbre aspects corresponding to specific instruments [33, 48]. Exploiting the availability of additional information such as musical scores, various score-informed source separation strategies have been proposed [19, 20, 21, 31, 71]. Having an explicit control over the various sources allows for building musically meaning equalizers (instead of simple frequency-based equalizer) that allow for amplifying or attenuating certain voices (instead of frequency bands), see, e. g., [38, 42] and Figure 10.

Decomposing a monaural audio signal into musical voices is, in general, an extremely difficult problem. A special case is the decomposition of a music signal into a harmonic and a percussive component. Here, various methods have been proposed based on matrix decompositions of a spectrogram representation using machine learning techniques [34, 23, 47, 64]. In [55], a simple and fast algorithm that does not require any training material is proposed. This iterative approach relies on the assumption that harmonic components correspond to horizontal and percussive components to vertical structures within a spectrogram.

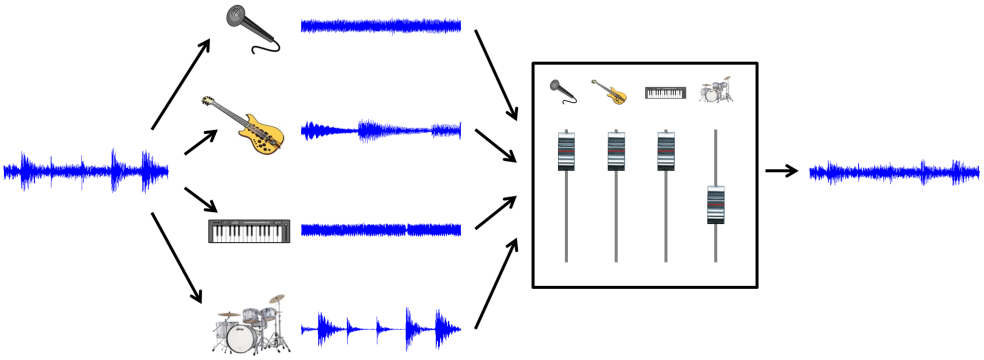


Figure 10 Instrument-wise equalization of a music recording (similar to [38, 42]).

In view of a sound track generation scenario, source separation and voice equalization techniques are important building blocks for the blending and morphing stage. Here, for example, one may want to suppress distracting voices or to amplify percussive components while concealing harmonic inconsistency. Actually, such techniques are also applied by DJs, who often amplify low-frequency bands while attenuating disturbing high-frequency bands in transition regions.

3.6 Indexing and Data Structures

In view of online capability of an overall sound track generation system, the *efficient* identification of suitable transition regions becomes an important issue. In the following, we want to touch on indexing and data structure issues.

Various indexing techniques have been applied for content-based audio retrieval. In case of audio identification, standard hashing techniques can be applied to obtain very efficient systems, see, e. g., [68]. For retrieval tasks on a lower specificity level, indexing become much harder because the *temporal order* of events, as also emphasized in [7], is of crucial importance for building up musically meaningful entities such as melodies or harmonic progressions. To account for temporal context, one often reverts to small chunks of audio also referred to as *audio shingles*, which leads, however, to features of high dimensionality. To index such high-dimensional shingles, techniques such as local sensitive hashing (LSH) have been applied for tasks such as *cover song identification* [6]. Here, being a document-based retrieval scenario, a bag-of-feature approach is applied with the features being the audio shingles. Such bag-of-feature approaches are not directly applicable to fragment-based retrieval scenarios such as audio matching. In [45], an indexing method is described based on inverted files which, however, only scales to medium size datasets. The idea of applying shingling and LSH-based indexing techniques to audio matching, where a single shingle corresponds to an entire audio clip of 10 to 20 seconds of duration, is investigated in [29].

Another idea to speed up the identification of transition candidates is to build up a graph-like data structure that explicitly encodes musical relations between audio clips. Such a data structure can be constructed from the given audio database in an off-line preprocessing step. As starting point, we want to take up an idea from the field of computer animation. Here, analogous to our music scenario, one important task consists in creating realistic, controllable motions from prerecorded motion capture sequences. In [44], a procedure is presented where a directed graph, referred to as *motion graph*, is constructed from a given

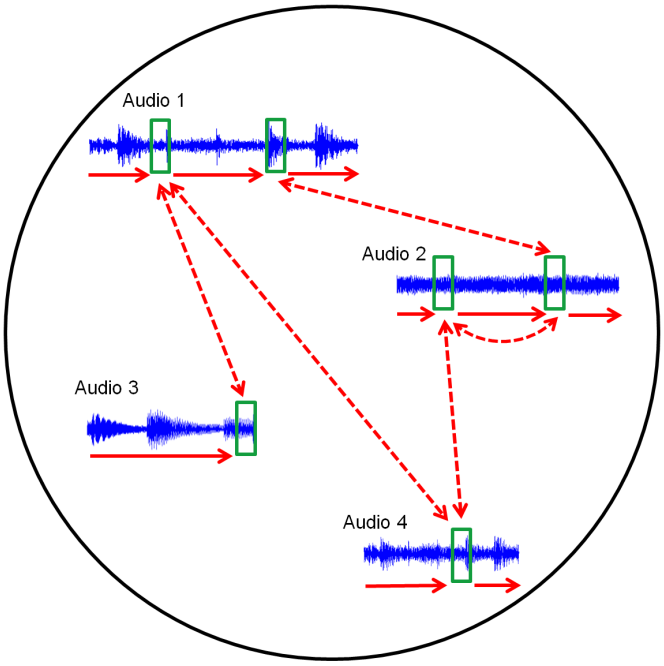


Figure 11 Music graph in analogy to the motion graph introduced in [44].

corpus of motion capture data. The edges of the graph contain either pieces of original motion data or automatically generated transitions, and the nodes serve as choice points where these small bits of motion join seamlessly. Motions can then be generated simply by building walks on the graph. Figure 11 illustrates this idea transferred to the music domain, see also [35] for a similar concept.

4 Conclusions and Future Work

The main goal of this contribution was to show how different aspects of music retrieval and audio processing come into play when dealing with applications such as data-driven sound track generation. Rather than presenting a concrete system, we sketched a possible pipeline for an online approach while discussing the necessary "ingredients" such as category-based music classification, content-based audio retrieval, beat tracking, time-scale modification, instrument equalization, and audio indexing. The intertwining and interaction of the various tasks is again summarized and illustrated by Figure 12. Each of the mentioned tasks constitutes itself a challenging research area with many open issues, in particular when dealing with various genre and styles of music—we have given numerous pointers to the literature that represent the state-of-the-art for the respective tasks.

Of course, when it comes to an actual realization and implementation of a concrete sound track generation system, many more challenges arise and a complete automatization of all steps neither seems feasible nor meaningful. However, there are many variants and more restricted sound track generation scenarios that come into reach. One such scenario is described in [69], where the duration of a given music recording is to be adjusted by suitably deleting, copying, and rearranging certain parts of the recording while keeping the flow of the music. Extending this scenario, a user may want to add background music to a slide show, where he specifies for each slide a desired music recording as well as a duration parameter.

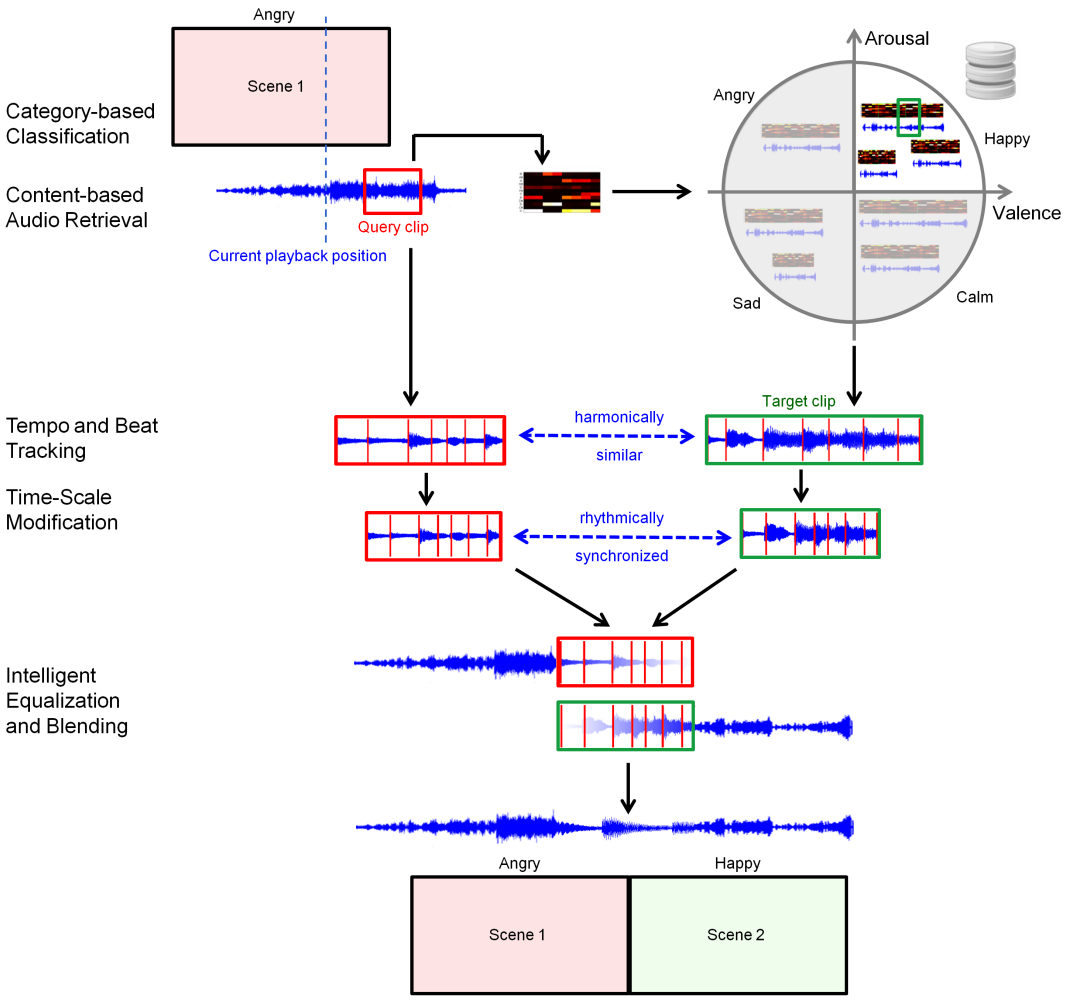


Figure 12 Possible pipeline for an automated sound track generation system.

Then, the task would be to automatically find and reassemble suitable parts of the recordings that not only fulfill the user constraints but also allow for euphonious transitions. Here, when the slide show is known in advance, an offline optimization procedure may be acceptable and efficiency issues become less significant. Furthermore, there may be different types of transitions a user may be interested in. For example, if there is a sudden event in the visual data stream, one may also want to have a surprising element in the sound track. Here, an abrupt change from one music clip to another may be acceptable or even desired. Instead of "complete solutions" that have been computed in a fully automated fashion, a user may rather need flexible tools that allow him to identify, modify, and assemble audio material in an intuitive and interactive way. Finally, perceptual issues need to be taken into account when it comes to the final assessment of the generated sound track. This itself constitutes an extremely difficult and interdisciplinary research area.

We hope that with this contribution we not only have given a useful overview of various tasks indicating challenges and future research directions, but could also give the reader an impression of the richness, depth and relevance of the research conducted in fields of music information retrieval and music processing.

5 Acknowledgment

This work has been supported by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) and the German Research Foundation (DFG MU 2682/5-1). We would like to express our gratitude to Frank Kurth and Hiromasa Fujihara for their helpful and constructive feedback.

References

1 Eric Allamanche, Jürgen Herre, Oliver Hellmuth, Bernhard Fröba, and Markus Cremer. AudioID: Towards content-based identification of audio material. In *Proceedings of the 110th AES Convention*, Amsterdam, NL, 2001.

2 Dan Barry, David Dorran, and Eugene Coyle. Time and pitch scale modification: A real-time framework and tutorial. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, 9 2008.

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

4 Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, 2005.

5 Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma. A review of algorithms for audio fingerprinting. In *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 169–173, St. Thomas, Virgin Islands, USA, 2002.

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

7 Michael Casey and Malcolm Slaney. The importance of sequences in musical similarity. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006.

8 Dave Cliff. Hang the DJ: Automatic sequencing and seamless mixing of dance-music tracks. Technical report, HP Laboratories Bristol, 2000.

9 Nick Collins. A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. In *AES Convention 118*, Barcelona, Spain, 2005.

10 Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, Elsevier, 2010.

11 David Cope. *Experiments in Musical Intelligence*. A-R Editions, Inc., 1996.

12 Matthew E.P. Davies and Mark D. Plumbley. Context-dependent beat tracking of musical audio. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1009–1020, 2007.

13 Mark Dolson. The phase vocoder: A tutorial. *Computer Music Journal*, 10(4):14–27, 1986.

14 Mark Dolson and Jean Laroche. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, 7(3):323–332, 1999.

15 David Dorran, Eugene Coyle, and Robert Lawlor. Audio time-scale modification using a hybrid time-frequency domain approach. In *Proceedings Workshop on Applications of Signal Processing (WASPAA)*, New Paltz, New York, USA, oct 2005.

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

17 Daniel P.W. Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007.

18 Daniel P.W. Ellis, Courtenay V. Cotton, and Michael I. Mandel. Cross-correlation of beat-synchronous representations for music similarity. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 57–60, Taipei, Taiwan, 2008.

19 Sebastian Ewert and Meinard Müller. Score-informed voice separation for piano recordings. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, pages 245–250, Miami, USA, 2011.

20 Sebastian Ewert and Meinard Müller. Using score-informed constraints for NMF-based source separation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

21 Joachim Ganseman, Paul Scheunders, Gautham J. Mysore, and Jonathan S. Abel. Source separation by score synthesis. In *Proceedings of the International Computer Music Conference (ICMC)*, New York, USA, 2010.

22 Loy Gareth. *Musimathics: The Mathematical Foundations of Music, Volume 1.* The MIT Press, 2006.

23 Olivier Gillet and Gaël Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech and Language Processing*, 16(3):529–540, March 2008.

24 Emilia Gómez. *Tonal Description of Music Audio Signals.* PhD thesis, UPF Barcelona, 2006.

25 Philippe Gournay, Roch Lefebvre, and Patrick-Andre Savard. Hybrid time-scale modification of audio. In *Audio Engineering Society Convention 122*, 5 2007.

26 F. Gouyon. *A computational approach to rhythm description: audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing.* PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2005. Available online: http://mtg.upf.edu/node/440.

27 Shahaf Grofit and Yizhar Lavner. Time-scale modification of audio signals using enhanced wsola with management of transients. *IEEE Transactions on Audio, Speech & Language Processing*, 16(1):106–115, 2008.

28 Peter Grosche and Meinard Müller. Extracting predominant local pulse information from music recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1688–1701, 2011.

29 Peter Grosche and Meinard Müller. Toward characteristic audio shingles for efficient cross-version music retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

30 Peter Grosche, Meinard Müller, and Craig Stuart Sapp. What makes beat tracking difficult? A case study on Chopin Mazurkas. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, pages 649–654, Utrecht, The Netherlands, 2010.

31 Yushen Han and Christopher Raphael. Desoloing monaural audio using mixture models. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 145–148, Vienna, Austria, 2007.

32 Gerhard Haupenthal. *Geschichte der Würfelmusik in Beispielen.* PhD thesis, Universität des Saarlandes, 1994.

33 Toni Heittola, Anssi Klapuri, and Tuomas Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 327–332, Kobe, Japan, 2009.

34 Marko Helen and Tuomas Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proc. EUSIPCO*, September 2005.

35 Alexander Höck, Frank Kurth, and Michael Clausen. Eine graphbasierte Indexstruktur zum inhaltsbasierten Audioretrieval. In *Proceedings of the Deutsche Jahrestagung für Akustik (DAGA)*, pages 185–186, Berlin, Germany, 2010.

36 X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. Ehmann. The 2007 mirex audio mood classification task: lessons learned. In *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, 2008.

37 Hiromi Ishizaki, Keiichiro Hoashi, and Yasuhiro Takishima. Full-automatic dj mixing system with optimal tempo adjustment based on measurement function of user discomfort. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 135–140, Kobe, Japan, 2009.

38 Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Instrument equalizer for query-by-example retrieval: Improving sound source separation based on integrated harmonic and inharmonic models. In *Proceedings of the International Conference for Music Information Retrieval (ISMIR)*, pages 133–138, Philadelphia, USA, 2008.

39 Tristan Jehan. *Creating Music by Listening*. PhD thesis, Massachusetts Institute of Technology, 2005.

40 Nicolas Juillerat, Stefan Mueller Arisona, and Simon Schubiger-Banz. A hybrid time and frequency domain audio pitch shifting algorithm. In *Audio Engineering Society Convention 125*, 10 2008.

41 Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. Music emotion recognition: a state-of-the-art review. In *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 255–266, 2010.

42 Tetsuro Kitahara, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Musical instrument recognizer "instrogram" and its application to music retrieval based on instrument similarity. In *IEEE international symposium on multimedia*, pages 265–272, San Diego, California, 2006.

43 Verena Konz, Meinard Müller, and Sebastian Ewert. A multi-perspective evaluation framework for chord recognition. In *Proceedings of the International Society for Music Information Retrieval Conference(ISMIR)*, pages 9–14, Utrecht, The Netherlands, 2010.

44 Lucas Kovar, Michael Gleicher, and Fréd éric Pighin. Motion graphs. *ACM Trans. Graph.*, 21(3):473–482, 2002.

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

46 Frank Kurth, Andreas Ribbrock, and Michael Clausen. Identification of highly distorted audio material for querying large scale data bases. In *Proceedings of the 112th AES Convention*, 2002.

47 D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Proc. NIPS*, pages 556–562, 2000.

48 Pierre Leveau, Emmanuel Vincent, Gaël Richard, and Laurent Daudet. Instrument-specific harmonic atoms for mid-level music representation. *IEEE Trans. Audio, Speech and Language Processing*, 16(1):116–128, 2008.

49 Heng-Li Lin, Yin-Tzu Lin, Ming-Chun Tien, and Ja-Ling Wu. Music paste: Concatenating music clips based on chroma and rythm features. In *10th International Society for Music Information Retreival Conference (ISMIR 2009)*, pages 213–218, 2009.

50 Riccardo Miotto and Nicola Orio. Automatic identification of music works through audio matching. In *ECDL*, pages 124–135, 2007.

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

52 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.

53 Meinard Müller and Sebastian Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the International Society for Music Information Retrieval Conference(ISMIR)*, pages 215–220, Miami, USA, 2011.

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

55 Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka, and Shigeki Sagayama. A real-time equalizer of harmonic and percussive components in music signals. In *Proc. ISMIR*, pages 139–144, Sepember 2008.

56 Richard Parncutt. A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, 11:409–464, 1994.

57 Geoffroy Peeters. Template-based estimation of time-varying tempo. *EURASIP Journal on Advances in Signal Processing*, 2007(1):158–158, 2007.

58 J. A. Russell. A circumspect model of affect. *Journal of Psychology and Social Psychology*, 39(6):1161, 1980.

59 N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, 2006.

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

61 Björn Schuller, Florian Eyben, and Gerhard Rigoll. Fast and robust meter and tempo recognition for the automatic discrimination of ballroom dance styles. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 217–220, 2007.

62 Emiru Tsunoo, Taichi Akase, Nobutaka Ono, and Shigeki Sagayama. Musical mood classification by rhythm and bass-line unit pattern analysis. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010.

63 G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. on Speech and Audio Processing*, 5(10):293–302, 2002.

64 Christian Uhle, Christian Dittmar, and Thomas Sporer. Extraction of drum tracks from polyphonic music using independent subspace analysis. In *Proc. ICA*, pages 843–847, April 2003.

65 Werner Verhelst and Marc Roelands. An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Minneapolis, USA, 1993.

66 David Vink. Adaptive music for video games. `http://www.gamecareerguide.com/features/768/student_thesis_adaptive_music_for_.php` (retrieved 19.02.2012), 2009.

67 Tuomas Virtanen. Unsupervised learning methods for source separation in monaural music signals. In Anssi P. Klapuri and Manuel Davy, editors, *Signal Processing Methods for Music Transcription*, chapter 6, pages 267–296. Springer US, 2006.

68 Avery Wang. An industrial strength audio search algorithm. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 7–13, Baltimore, USA, 2003.

69 Stephan Wenger and Marcus Magnor. Constrained example-based audio synthesis. In *Proceedings of the 2011 IEEE International Conference on Multimedia and Expo (ICME 2011)*, Barcelona, Spain, July 2011.

70 Guy Whitmore. Design with music in mind: A guide to adaptive audio for game designers. `http://www.gamasutra.com/resource_guide/20030528/whitmore_01.shtml` (retrieved 19.02.2012), 2003.

71 John Woodruff, Bryan Pardo, and Roger B. Dannenberg. Remixing stereo music with score-informed source separation. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 314–319, 2006.

72 Ruohua Zhou, Marco Mattavelli, and Giorgio Zoia. Music onset detection based on resonator time frequency image. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1685–1695, 2008.

Music Information Retrieval: An Inspirational Guide to Transfer from Related Disciplines

Felix Weninger[*1], Björn Schuller[1], Cynthia C. S. Liem[†2], Frank Kurth[3], and Alan Hanjalic[2]

1 **Technische Universität München**
Arcisstraße 21, 80333 München, Germany
weninger@tum.de

2 **Delft University of Technology**
Mekelweg 4, 2628 CD Delft, The Netherlands
c.c.s.liem@tudelft.nl

3 **Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie FKIE**
Neuenahrer Straße 20, 53343 Wachtberg, Germany
frank.kurth@fkie.fraunhofer.de

Abstract

The emerging field of Music Information Retrieval (MIR) has been influenced by neighboring domains in signal processing and machine learning, including automatic speech recognition, image processing and text information retrieval. In this contribution, we start with concrete examples for methodology transfer between speech and music processing, oriented on the building blocks of pattern recognition: preprocessing, feature extraction, and classification/decoding. We then assume a higher level viewpoint when describing sources of mutual inspiration derived from text and image information retrieval. We conclude that dealing with the peculiarities of music in MIR research has contributed to advancing the state-of-the-art in other fields, and that many future challenges in MIR are strikingly similar to those that other research areas have been facing.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems, I.5 Pattern Recognition

Keywords and phrases Feature extraction, machine learning, multimodal fusion, evaluation, human factors, cross-domain methodology transfer

Digital Object Identifier 10.4230/DFU.Vol3.11041.195

1 Introduction

Music Information Retrieval (MIR) still is a relatively young field: Its first dedicated symposium, ISMIR, was held in 2000, and a formal society for practitioners in the field, taking over the ISMIR acronym, was only established in 2008. This does not mean that all work in MIR needs to be newly invented: Analogous or very similar topics and areas to those currently of interest in MIR research may already have been researched for years, or even decades, in neighboring fields. Reusing and transferring findings from neighboring fields, MIR research can jump-start and stand on the shoulders of giants. At the same time, the

* Felix Weninger is funded by the German Research Foundation through grant no. SCHU 2508/2-1.
† The work of Cynthia Liem is supported in part by the Google European Doctoral Fellowship in Multimedia.

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 195–216

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

nature of music data may pose constraints or peculiarities that press for solutions beyond the trodden paths in MIR, and thus can be of inspiration the other way around too. Such opportunities for methodology transfer, both to and from the MIR field, are the focus of this chapter.

In engineering contexts, audio typically is considered to be the main modality of music. From this perspective, an obvious neighboring field to look at is automatic speech recognition (ASR), which just like MIR strives to extract information from audio signals. Section 2 will discuss several methodology transfers from ASR to MIR, while Section 3 gives a detailed example of one of the first successful transfers from MIR back to ASR. Section 4 focuses on the topic of evaluation, in which current MIR practice has strong connections to classical approaches in Text Information Retrieval (IR). Finally, in Section 5, we consider MIR from a higher-level, more philosophical viewpoint, pointing out similarities in open challenges between MIR and Content-Based Image and Multimedia Retrieval, and arguing that MIR may be the field that can give a considerable push towards addressing these challenges.

2 Synergies between Speech and Music Analysis

As stated above, it is hardly surprising that audio-based MIR has been influenced by ASR research—as obvious opportunities to transfer ASR technologies to MIR, lyrics transcription [38] or keyword spotting in lyrics [17] can be named. Yet, there are more intrinsic synergies between speech and music analysis, where similar methodologies can be applied to seemingly different tasks. These will be the focus of the following section. We point out areas where speech and music analysis have been sources of mutual inspiration in the past, and sketch some opportunities for future methodology transfer.

2.1 Multi-Source Audio Analysis in Speech and Music

Generally, music signals are composed of multiple sources, which can correspond to instruments, singer(s), or the voices in a polyphonic piano piece; thus, aspects of multi-source signal processing can be considered as an integral part of MIR. Similarly, research on speech recognition in the presence of interfering sources (environmental noise, or even other speakers) has a long tradition, resulting in numerous studies on source separation and model-based robust speech recognition. Many approaches for speech source separation deal with multi-channel input from microphone arrays by beamforming, i. e., exploitation of spatial information. An example of such beamforming in music signals is the well-known 'karaoke effect' to remove the singing voice in commercial stereophonic recordings: Many popular songs are mixed with the vocals being equally distributed to the left and right channels, which corresponds to a center position of the the vocalist in the recording/playback environment. In that case, the vocals can be simply eliminated by channel subtraction, which can be regarded as a trivial example of integrating spatial information into source separation. However, to highlight the aspects of methodology transfer, we restrict the following discussion to monaural (single-channel) analysis methods: We argue that the constraints of music signal processing—where usually no more than two input channels are available—have leveraged a great deal of research on monaural source separation, which has been fruitful for speech signal processing in turn. In this section, we attempt a unified view on monaural audio source separation in speech and music, presenting a rough taxonomy of tasks and applications where synergies are evident. This taxonomy is oriented on the general procedure depicted in Figure 1, depending on which of the system components (source models, transcription/alignment, synthesis) are present.

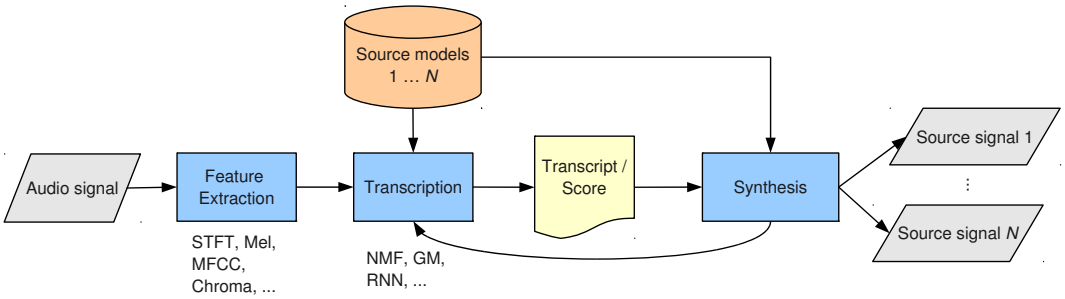


Figure 1 A unified view on monaural multi-source analysis of speech and music. Spectral (short-time Fourier Transform, STFT) or cepstral features (MFCCs) are extracted from the audio signal, yielding a transcription based on non-negative matrix factorization (NMF), graphical models (GM), recurrent neural networks (RNN) or other machine learning algorithms. The transcription can be used to synthesize signals corresponding to the sources or to enable (more robust) transcription in turn.

Polyphonic transcription and multi-source decoding

The goal of these tasks is not primarily the synthesis of each source as a waveform signal, but to gain a higher-level transcription of each source's contributions, e. g., the notes played by different instruments, or the transcription of the utterances by several speakers in a cross-talk scenario (the 'cocktail party problem'). Polyphonic transcription of monaural music signals can be achieved by sparse coding through non-negative matrix factorization (NMF) [64, 68], representing the spectrogram as the product of note spectra and a sparse non-negative activation matrix. These sparse NMF techniques have successfully been ported to the speech domain to reveal the phonetic content of utterances spoken in multi-source environments [18]: Determining the individual notes played by various instruments and their position in the spectrogram can be regarded as analogous to detecting individual phonemes in the presence of interfering talkers or environmental noise. An important common feature of these 'joint decoding' approaches for multi-source speech and music signals is the explicit modeling of parallel occurrence of sources; this can also be done by a graphical model representation of probabilistic dependencies between sources, as demonstrated in [69] for multi-talker ASR. Furthermore, polyphonic transcription approaches that use discriminative models for multiple note targets [46] or one-versus-all classification [50] seem to be partly inspired by 'multi-condition training' in ASR, where speech overlaid with interfering sources is presented to the system in the training stage, to learn to recognize speech in the presence of other sources. Finally, to contrast transcription or joint decoding approaches to the methods presented in the remainder of this section, we note that the former can principally be used to resynthesize signals corresponding to each of the sources [69], yet this is not their primary design goal; results are sometimes inferior to dedicated source separation approaches [19, 73].

Leading voice extraction and noise cancellation

For many MIR applications, the leading voice is of particular relevance, e. g., the voice of the singer in a karaoke application. Similarly, in many speech-based human-human and human-computer interaction scenarios, including automatic analysis of meetings, voice search or mobile telephony, the extraction of the primary speech source, which delivers the relevant content, is sufficient. This application requires modeling of the characteristics of the primary source, and speech and music processing considerably differ in this respect; unifying the

approaches will be an interesting question for future research. In music signal processing, main melody extraction is often related to predominance: It is assumed that the singing voice contributes the most to the signal energy[1]. Thus, extraction of the leading voice can be achieved with little explicit knowledge, e. g., by fixing a basis of sung notes and estimating the vocal tract impulse response in an extension of NMF to a source-filter model [14]. In speech processing, one usually does not rely on the assumption that the wanted speech is predominant in a recording, as signal-to-noise ratios can be negative in many realistic scenarios [9]. Hence, one extends the previous approaches by rather precise modeling of speech, often in a speaker-dependent scenario. Still, combining knowledge about the spectral characteristics of the speech with unsupervised estimation of the noise signal, in analogy to the unsupervised estimation of the accompaniment in [14], results in a semi-supervised approach for speech extraction as, e. g., in [48]. In contrast, often a pre-defined model for the background such as in [19, 53, 73] is used in a supervised source separation framework, and this kind of background modeling can be applied to leading voice extraction as well: Assuming the characteristics of the instrumental accompaniment of the singer are similar in vocal and non-vocal parts, a model of the accompaniment can be built; this allows estimating the contribution of the singing voice through semi-supervised NMF [21].

Instrument Separation and the Cocktail Party Problem

As laid out above, leading voice extraction or speech enhancement can be conceived as source separation problems with two sources. A generalization of this problem to extraction of multiple sources, or sources with large spectral similarity such as in instrument separation or the 'cocktail party' scenario, from a monophonic recording generally requires more complex source modeling. This can include temporal dependencies: In [45], NMF is extended to a non-negative Hidden Markov Model for extraction of the individual speakers from a multi-talker recording. Including temporal dependencies appears promising for music contexts as well, e. g., for separation of (repetitive) percussive and (non-repetitive) harmonic sources; furthermore, this approach is purely data-based and generalizes well to multiple sources.

In music signal processing, especially for classical music, higher-level knowledge can be incorporated into signal separation by means of score information (score-informed source separation) [15, 24]. Not only does this allow to cope with large spectral similarity, but it also enables separation by semantic aspects, which would be infeasible from an acoustic feature representation, and/or allows for user guidance; for instance, the passages played by the left and right hand in a piano recording can be retrieved [15]. Transferring this approach to the speech domain, we argue that while in most speech-related applications availability of a 'score' (i. e., a ground truth speaker diarization including overlap and transcription) cannot be assumed, score-informed separation techniques could be an inspiration to built iterative, self-improving methods for cross-talk separation, speech enhancement and ASR, recognizing what has been said by whom and exploiting that higher-level knowledge in the enhancement algorithm.

2.2 Combined Acoustic and Language Modeling

Language modeling techniques are found in MIR, e. g., to model chord progressions [47, 58, 80] or playlists [36]. Conversely, the prevalent usage of language models in ASR is

[1] Other common assumptions are that the singing voice is the highest voice among all instruments, or that it is characterized by vibrato.

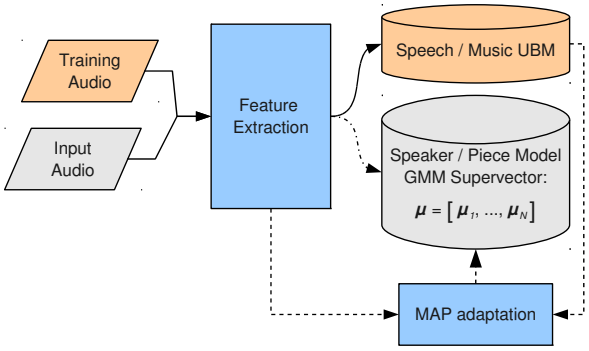


Figure 2 Use of universal background models (UBM) in speech and music processing: A generic speech/music model (UBM) is created from training audio. A speaker/piece model can be generated directly from training audio (dashed-dotted curve) or from the UBM by MAP adaptation (dashed lines). In the latter case, the parameters of the adapted model (e. g., the mean vector μ in case of GM modeling) yield a fingerprint (*supervector*) of the speaker or the music piece.

to calculate combined acoustic-linguistic likelihoods for speech decoding: Informally, the acoustic likelihood of a phoneme in an utterance is multiplied with a language model likelihood of possible words containing the phoneme to integrate knowledge about word usage frequencies (unigram probabilities) and temporal dependencies (n-grams) [82]. This immediately translates to chord recognition: For instance, unigram probabilities can model the fact that major and minor chords are most frequent in Western music, and there exist typical chord progressions that can be modeled by n-grams [56]. Thus, accuracy of chord recognition can be improved by combined acoustic and language modeling in analogy to ASR [8, 29]. A different approach to combined acoustic and language modeling is taken in [30] for genre classification: Music is encoded in a symbolic representation derived from clustered acoustic features, which is then encoded in a language model for different genres.

2.3 Universal Background Models in Speech Analysis and Music Retrieval

Recent developments in content-based music retrieval include methodologies that were introduced for speaker recognition and verification. These include universal background models (UBM)—trained from large amounts of data, and representing generic speech as opposed to the speech characteristics of an individual—and Gaussian Mixture Model (GMM) supervectors [4, 35, 81]. GMM supervectors are equivalent to the parameters of a Gaussian Mixture UBM adapted to the speech of a single speaker (usually only few utterances). Hence, they allow for effective and efficient computation of a person's speech 'fingerprint', i. e., its representation in a concise feature space suitable for a discriminative classifier. The generic approaches incorporating UBMs for speech and music classification are shown in Figure 2: A basic speaker verification algorithm uses a UBM to represent the acoustic parameters of a large set of speakers, while the speaker to be verified is modeled with a specialized GMM. For an utterance to be verified, a likelihood ratio test is conducted to determine whether the speaker model delivers sufficiently higher likelihood than the UBM. Translating this paradigm to music retrieval, one can cope with out-of-set events—i. e., that the user may be querying for a musical piece not contained in the database. Specific pieces in the database are represented ('fingerprinted') by Gaussian mixture modeling of acoustic features, while the UBM is a generic model of music. Then, the likelihoods of the query under the specialized GMMs versus the UBM allow out-of-set classification [39].

On the other hand, adapting the UBM to a specific music piece using maximum-a-posteriori (MAP) adaptation yields an audio fingerprint in shape of the adapted model's mean (and possibly variance) vectors. These fingerprints can be classified by discriminative models such as Support Vector Machines (SVMs), resulting in the GMM-SVM paradigm which has become standard in speaker recognition in the last years. In [5], the GMM-SVM approach was successfully applied to music tagging in the 2009 MIREX evaluation; recent studies [6, 7] underline the suitability of the approach to analyze music similarity for recommender systems.

2.4 Transfer from Paralinguistic Analysis

To elucidate a further opportunity for methodology transfer from the speech domain, we consider the field of paralinguistic analysis (i. e., retrieving other information from speech beyond the spoken text), which is believed to be important for natural human-machine and computer mediated human-human communication. Particularly, we address synergies between speech emotion recognition and music mood analysis: While relating to different concepts of emotion (or mood), the overlap in the methodologies and the research challenges are striking. At first, we would like to recall the subtle difference between those fields: Speech emotion recognition aims to determine the emotion of the speaker, which is—for most practical applications such as in dialog systems—the emotion perceived by the conversation partner; conversely, music mood analysis does not primarily assess the (perceived) mood of the singer, but rather the overall perceived mood in a musical piece—often, that is the intended mood, i. e., the mood as intended by the composer (or songwriter). Despite these differences, in the result, similar pattern recognition techniques have been proven useful in practice.

For instance, in order to assess the emotion of a speaker, combining 'what' is said with 'how' it is said, i. e., fusing acoustic with linguistic information, has been shown to increase robustness [78]—and similar results have been obtained in music mood analysis when considering lyrics and audio features [26, 57]. Apart from low-level acoustic and linguistic features, specific music features seem to contribute to music mood perception, and hence, recognition performance, including the harmonic 'language' (chord progression) and rhythmic structure [60], which necessitates efficient fusion methods as, e. g., for audio-visual emotion recognition. Besides, similarly to emotion in speech [77], music mood classification is lately often turned into a regression problem [60, 79] in target dimensions such as the arousal-valence plane [55], in order to avoid ambiguities in categorical 'tags' and improve model generalization.

Furthermore, when facing real-life applications, the issue of non-prototypical instances—i. e., musical pieces that are not pre-selected by experts as being representative for a certain mood—has to be addressed: It can be argued that a recommender system based on music mood should retrieve instances associated with high degrees of, e. g., happiness or relaxation from a large music archive. Here, music mood recognition can profit from the speech domain as this task bears some similarity to applications of speech emotion recognition such as anger detection, where emotional utterances have to be discriminated from a vast amount of neutral speech [66]. Relatedly, whenever instances to be annotated with the associated mood are not pre-selected by experts according to their prototypicality, the establishment of a robust ground truth, i. e., consistent assessment of the music mood by multiple human annotators, becomes non-trivial [27]. This might foster the development of quality control and 'noise cancellation' methods for subjective music mood ratings [60], as developed for speech emotion [20], in the future.

Finally, in the future, we might see a shift towards recognizing the affective state of singers themselves: First attempts have been made to estimate the 'enthusiasm' of the singer [10], which is arguably positively correlated with both arousal and valence; hence, the task is somewhat similar to recognition of level of interest from speech as in [78]. Another promising research direction might be to investigate long-term singer traits instead of short-term states such as emotion: Such traits include age, gender [59], body shape and race, all of which are known to be correlated with acoustic parameters, and can be useful in category-based music retrieval or identifying artists from a meta-database [74]. In a similar vein, the analysis of voice quality and 'likability' [72] could be a valuable source of inspiration for research on synthesis of singing voices.

3 From Music IR to Speech IR: An Example

Starting from the general overview above, we now discuss a particular example on how technologies from both domains of music and speech IR interact with each other. In particular, we start with the well known MFCC (Mel Frequency Cepstral Coefficients) features from the speech domain which are used to analyze signals based on an auditory filterbank. This results in representing a speech signal by a temporal feature sequence correlating with certain properties of the speech signal. We then review corresponding music features and their properties, with a particular interest on representing the harmonic progression of a piece of music using chroma-type features. This, in turn, inspires a class of speech features correlating with the phonetic progression of speech.

Concerning possible applications, chroma-type features can be used to identify fragments of audio as being part of a musical work regardless of the particular interpretation. Having sketched a suitable matching technique, we subsequently show how similar techniques can be applied in the speech domain for the task of keyphrase spotting.

Whereas the latter matching techniques focus on local temporal regions of audio, more global properties can be analyzed using self-similarity matrices. In music, such matrices can be used to derive the general repetitive structure (related to the musical form) of an audio recording. When dealing with two different interpretations of a piece of music, such matrices can be used to derive a temporal alignment between the two versions. We discuss possible analogies in speech processing and sketch an alternative approach to text-to-speech alignment.

3.1 Feature Extraction

Many audio features are based on analyzing the spectral contents of subsequent short temporal segments of a target signal by using either a Fourier transform or a filter-bank. The resulting sequence of vectors is then further processed depending on the application. As an example, the popular MFCC features which have been successfully applied in automatic speech recognition (ASR) are obtained by applying an auditory filterbank based on log-scale center frequencies, followed by converting subband energies to a dB- (log-) scale, and applying a discrete cosine transform [51]. The logarithmic compression in both frequency and signal power serves to weight the importance of events in both domains in a way a human perceives them. Because of their ability to describe a short-time spectral envelope of an audio signal in a compact form, MFCCs have been successfully applied to various speech processing problems apart from ASR, such as keyword spotting and speaker recognition [54]. Also in Music IR, MFCCs have been widely used, e. g., for representing the timbre of musical instruments or speech-music discrimination [34].

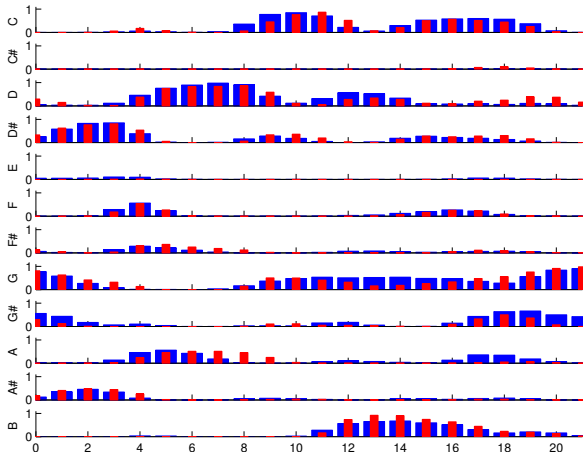


Figure 3 Chroma-based CENS features obtained from the first measures (20 seconds) of Beethoven's 5th Symphony in two interpretations by Bernstein (blue) and Sawallisch (red).

While MFCCs are mainly motivated by auditory perception, music analysis is frequently performed based on features motivated by the process of sound generation. Chroma features for example, which have received an increasing amount of attention during the last ten years [2], rely on the fixed frequency (semitone) scale as used in Western music. To obtain a chroma feature for a short segment of audio, a Fourier transform of that segment is performed. Subsequently, the spectral coefficients corresponding to each of the twelve musical pitch classes (the *chroma*) C, C$^\sharp$, D,..., B are individually summed up to yield a 12-dimensional chroma vector. In terms of a filterbank, this process can be seen as applying octave-spaced comb-filters for each chroma.

From their construction, chroma features do well-represent the local harmonic content of a segment of music. To describe the temporal harmonic progression of a piece of music, it is beneficial to combine sequences of successive chroma features to form a new feature type. CENS-features (chroma energy normalized statistics) [43] follow this approach and involve calculating certain short-time statistics on the chroma features' behaviour in time, frequency, and energy. By adjusting the temporal size of the statistics window, CENS-feature sequences of different temporal resolutions may be derived from an input signal. Figure 3 shows the resulting CENS feature sequences derived from two performances of Beethoven's 5th Symphony.

In the speech domain, a possible analogy to the local harmonic progression of a piece of music is the phonetic progression of a spoken sequence of words (a *phrase*). To model such phonetic progressions, the concept of energy normalized statistics (ENS) has been transferred to speech features [70]. This approach uses a modified version of MFCCs, called HFCCs (human factor cepstral coefficients), where the widths of the mel-spaced filter bands are chosen according to the bark scale of critical bands. After applying the above statistics computations, the resuling features are called HFCC-ENS. Figure 6 (c) and (d) show sequences of HFCC-ENS features for two spoken versions of the same phrase. Experiments show that due to the process of calculating statistics, HFCC-ENS features are better adapted to the phonetic progression in speech than MFCCs [70].

3.2 Matching Techniques

In this section, we describe some matching techniques that use audio features in order to automatically recognize audio signals. Current approaches to ASR or keyword spotting employ suitable HMMs trained to individual words (or subword entities) to be recognized. Usually, speaker-dependent training results in a significant improvement in recognition rates and accuracy. Older approaches used dynamic time warping (DTW) which is simpler to implement and bears the advantage of not requiring prior training. However, as the flexibility of DTW in modeling speech properties is restricted, it is not as widely used in applications as HMMs are [52]. In the context of music retrieval, DTW and variants thereof have, however, regained considerable attention [40].

As particular example, we consider the task of audio matching: Given a short fragment of a piece of audio, the goal is to identify the underlying musical work. A refined task would be to additionally determine the position of the given fragment within the musical work. This task can be cast into a database search: given a short audio fragment (the *query*) and a collection of "known" pieces of music (the *database*), determine the piece in the database the query is contained in (the *match*). Here a restricted task, widely known as *audio identification*, only reports a match if the query and a match correspond to the same audio recording [1, 71]. In general audio matching, however, a match is also reported if a query and the database recording are different performances of the same piece of music. Whereas audio identification can be very efficiently performed using low-level features describing the physical waveform, audio matching has to use more abstract features in order to identify different interpretations of the same musical work. In Western classical music, different interpretations can exhibit significant differences, e. g., regarding tempo and instrumentation. In popular music, different interpretations include cover songs that may exhibit changes in musical style as well as mixing with other audio sources [62].

The introduced CENS features are particularly suitable to perform audio matching for music that possess characteristic harmonic progressions. In a basic approach [43], the query and database signals are converted to feature sequences $q = (q_1, \ldots, q_M)$ and $d = (d_1, \ldots, d_N)$, where each of the q_i and d_j are 12-dimensional CENS vectors. Matching is then performed using a cross-correlation like approach, where a similarity function $\Delta(n) := \frac{1}{M} \sum_{\ell=1}^{M} \langle q_\ell, d_{n-1+\ell} \rangle$ gives the similarity of query and database at position n. Using normalized feature vectors, values of Δ in a range of $[0, 1]$ can be enforced. Figure 4 (top) shows an example of a resulting Δ when using the first 20 seconds of the Bernstein interpretation (see Figure 3) as a query to a database containing, among other material, two different versions of Beethovens Fifth by Bernstein and Sawallisch respectively. Positions corresponding to the seven best matches are indicated in green. The first six matches correspond to the three occurrences of the query (corresponding to the famous theme) within the two performances. Tolerance with respect to different global tempi may be obtained in two ways: On the one hand, one may calculate p time-scaled versions of the feature sequence q by simply changing the statistics parameters (particularly window size and sampling rate) during extraction of the CENS features. This process is then followed by p different evaluations of Δ. On the other hand, the correlation-based approach to calculate a cost function may be replaced by a variant of subsequence DTW. Experiments show that both variants perform comparably.

Coming back to the speech domain, the some audio matching approach can be applied to detect short sequences of words or *phrases* within a speech recording. Compared to classical keyword spotting [28, 76], this kind of *keyphrase* spotting is particularly beneficial when the target phrase consists of at least 3-4 words [70]. Advantages inherited from using the above HFCC-ENS features for this task are speaker and also gender independence. More important,

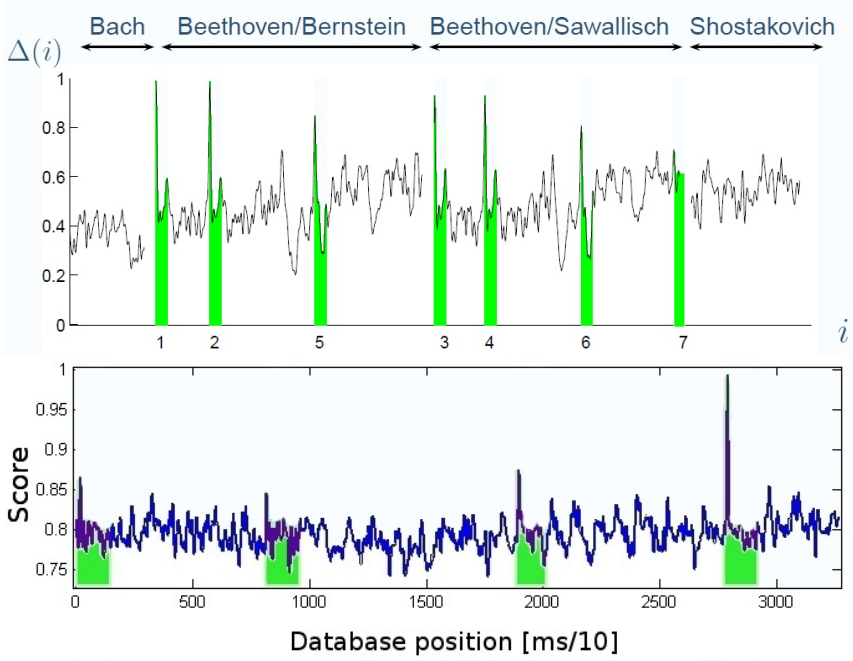


Figure 4 Top: Similarity function Δ obtained in scenarios of audio matching for music. Bottom: Similarity function Δ obtained in keyphrase matching.

no prior training is required which makes this form of keyphrase spotting attractive for scenarios with sparse resources. Figure 4 (bottom) shows an example where the German phrase *"Heute ist schönes Frühlingswetter"* was used as a query to a database containing a total of 40 phrases spoken by different speakers. Among those are four versions of the query phrase each by a different speaker. All of them are identified as matches (indicated in green) by applying a suitable peak picking strategy on the similarity function.

3.3 Similarity Matrices: Synchronization and Structure Extraction

To obtain the similarity of a query q and a particular position of a database document d, a similarity function Δ has been constructed by averaging M local comparisons $\langle q_i, d_j \rangle$ of features vectors q_i and d_j. In general, the similarity between two feature sequences $a = (a_1, \ldots, a_K)$ and $b = (b_1, \ldots, b_L)$ can be characterized by calculating a *similarity matrix* $S_{a,b} := (\langle a_i, b_j \rangle)_{1 \leq i \leq K, 1 \leq j \leq L}$ consisting of all pair-wise comparisons. Figure 5 (left) shows an example of a similarity matrix. Color coding is chosen in a way such that dark regions indicate a high local similarity and light regions correspond to a low local similarity. The diagonal-like trajectory running from the lower left to the upper right thus expresses the difference in the local tempo between the two underlying performances.

Based on such trajectories, similarity matrices can be used to temporally *synchronize* musically corresponding positions of the two different interpretations [25, 44]. Technically, this amounts to finding a *warping path* $p := (x_i, y_i)_{i=1}^{P}$ through the matrix, such that $\delta(p) := \sum_{i=1}^{P} \langle a_{x_i}, b_{y_i} \rangle$ is minimized. Warping paths are restricted to start in the lower left corner, $(x_1, y_1) = (1, 1)$, end in the upper right, $(x_P, y_P) = (K, L)$, and obey certain step conditions, $(x_{i+1}, y_{i+1}) = (x_i, y_i) + \sigma$. Two frequently used step conditions are $\sigma \in \{(0, 1), (1, 0), (1, 1)\}$ and $\sigma \in \{(2, 1), (1, 2), (1, 1)\}$. In Figure 5 (left) a calculated warping path is indicated in red color.

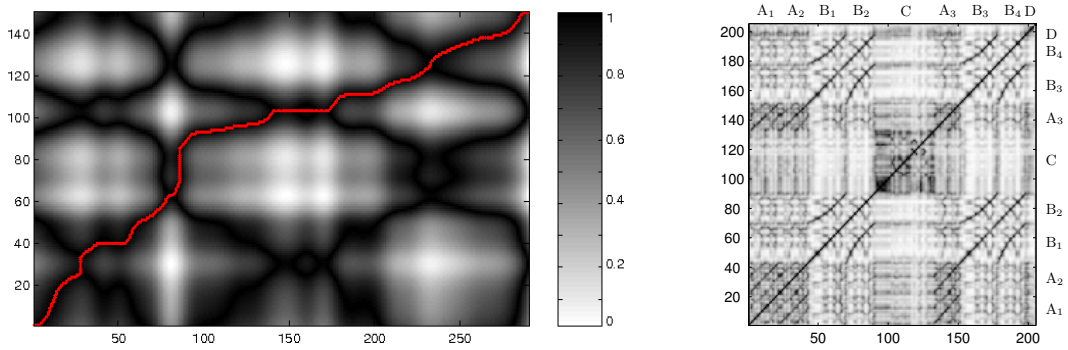


Figure 5 Left: Example of a similarity matrix with warping path indicated in red color. **Right:** Self-similarity matrix for a version of Brahms Hungarian Dances no. 5. The extracted musical structure $A_1A_2B_2CA_3B_3B_4D$ is indicated. (Figures from [40].)

Besides synchronizing two audio recordings of the same piece, the latter methods can be used to time-align musically corresponding events *across* different representations. As a first example, consider a (symbolic) MIDI representations of the piece of music. In a straightforward approach, an audio version of the MIDI can be created using a synthesizer. Then, CENS features are obtained from the synthesized signal, thus allowing a subsequent synchronization with another audio recording (in this context an audio recording obtained from a real performance). Alternatively, CENS features may be generated *directly* from the MIDI [25]. In a second example, scanned sheets of music (i. e., digital images) can be synchronized to audio recordings, by first performing optical music recognition (OMR) on the scanned images, producing a symbolic, MIDI-like, representation. In a second step, the symbolic representation is then synchronized to the audio recording as described before [16]. This process is illustrated in Figure 6 (left). Besides the illustrated task of audio synchronization, the automatic alignment of audio and lyrics has also been studied [37], suggesting the usability of synchronization techniques for human speech.

Transfered to the speech domain, such synchronization techniques can be used to time-align speech signals with a corresponding textual transcript. Similarly to using a music synthesizer on MIDI input to generate a music signal, a text-to-speech (TTS) system can be used to create a speech signal. Subsequently, DTW-based synchronization can be performed on HFCC-ENS feature sequences extracted from both speech signals [11], see Figure 6 (right).

Text-to-speech synchronization as decribed here may be applied for example to political speeches or audio books. We note that a more classical way of performing this synchronization consists of first performing ASR on the speech signal, resulting in an approximate textual transcript. In a second step, both transcripts can then by synchronized by suitable text-based DTW techniques [23].

ASR-based synchronization is advantageous in case of relatively good speech quality or when a prior training to the speaker is possible. In this case, the textual transcript will be of sufficiently high quality and a precise synchronization is possible. Due to the smoothing process involved in the ENS calculation, TTS-based synchronization typically has a lower temporal resolution which has an impact on the synchronization accuracy. However, in scenarios with a high likelihood of ASR-errors, TTS-based synchronization can be beneficial.

Variants of the DTW-based music synchronization perform well if the musical structure underlying a and b are the same. In case of structural differences, advanced synchronization methods have to be used [41]. To analyze the structure of a music signal, the *self-similarity matrix* $S_a := S_{a,a}$ of the corresponding feature sequence a can be employed. As an example,

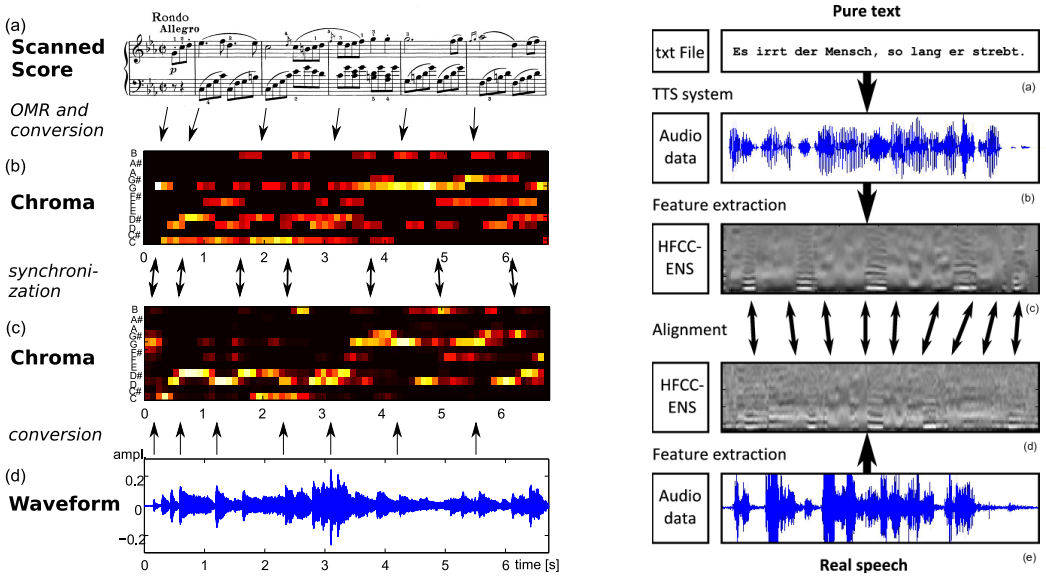


Figure 6 Left: Score-Sheet to audio synchronization—(a) Score fragment, (b) Synthesized Chroma features, (c) Chroma obtained from audio recording (d). **Right:** Text to audio synchronization—(a) Text, (b) Synthesized speech, (c) HFCC-ENS features of synthesized speech, (d) HFCC-ENS features of natural speech (e).

Figure 5 depicts the self-similarity matrix of an interpretation of Brahms Hungarian Dances no. 5 by Ormandy. Darker trajectories on the side diagonals indicate repeating music passages. Extraction of all such repetitions and systematic structuring can be used to deduce the underlying musical form. In our example, the musical form $A_1A_2B_2CA_3B_3B_4D$ is obtained by following an approach to calculate a complete list of all repetitions [42].

Concluding, we discuss possible applications of structure analysis in the speech domain, where one first has to ask for suitable analogies of *structured speech*. In contrast to music analysis, where the target signal to be analyzed frequently corresponds to a complete piece of music, in speech one frequently analyses unstructured speech fragments such as isolated sequences of sentences or a dialog between two persons. Lower-level examples of speech structure relevant for unstructured speech could be repeated words, phrases, or sentences. More structure on a higher level could be expected from speech recorded in special contexts such as TV shows, news, phone calls, or radio communication. An even closer analogy to music analysis could be the analysis of recited poetry.

4 Evaluation: The Information Retrieval Legacy

We now move on to another field with considerable influences on MIR research: Information Retrieval (IR). This field, after which the MIR field was named, deals with storing, extracting and retrieving information from text documents. The information can be both syntactic and semantic, and topics of interest cover a wide range, involving feature representations, full database systems, and information-seeking behavior of users.

Evaluation in MIR work, especially in retrieval settings, has largely been influenced by IR evaluation, with *Precision*, *Recall* and the *F-measure* as most stereotypical evaluation criteria. However, already in the first years of the MIR community benchmark evaluation endeavor, the Music Information Retrieval EXchange (MIREX), the need arose to find

significance levels for system results. Earlier findings from the Text REtrieval Conference (TREC) benchmarking efforts led to the adoption of Friedman's ANOVA with Tukey-Kramer "Honestly Significant Difference" post-hoc correction [13], which subsequently were widely adopted in the presentation of MIREX results.

Not all of the IR practices were immediately transferable to MIR evaluation: many MIREX tasks turned out to be specialized enough to a degree that they require task-specific evaluation criteria. In addition, precision and recall have frequently been challenged for their appropriateness. In cover song retrieval and audio matching settings, recall may be the most appropriate, since the goal would be to retrieve as many matching items or fragments as possible [61]. On the other hand, in web-scale environments, the amount of data will be so huge that striving for recall will not make sense anymore. In addition, in multimedia settings one can wonder if precision would be an appropriate measure at all, since user data suggests that multimedia search is more of an entertaining browsing activity rather than a focused information need with a concrete query and an establishable ground truth [63]. Exactly the same will hold for music search.

Nonetheless, there still are existing IR evaluation findings that provide useful opportunities for strengthening evaluation in MIR, an important area being that of *meta-evaluation* [67]. Through meta-evaluation, the experimental validity of (M)IR experiments can be assessed. This validity can be assessed according to different subcategories, which are listed below together with reflections on the way in which they are applicable to the MIR domain:

Construct validity

The extent to which the variables of an experiment correspond to the theoretical meaning of the concept they are intended to measure. To give an example for MIR, it is tempting to try to infer music 'mood' from features present in musical audio (e.g. presence of major/minor chords and tonalities); however, the situation is often more complicated. Most importantly, mood implies a human property, and is usually experienced due to a certain (multimodal) context. Thus, in order to truly address mood, work related to music and mood should not only look at audio features and take the user and this context into account.

Content validity

The extent to which the experimental units reflect and represent the elements of the domain under study. For example, an experiment aimed at measuring 'audio similarity' between songs cannot be (solely) based on item co-occurrences of these songs in a social network.

Convergent validity

The extent to which the results of an experiment agree with other results they should be related with (both theoretical and experimental). As an example from the MIR domain, a good tempo estimator should involve a good beat estimating component. Thus, this beat estimating component would be expected to perform well on beat extraction tasks.

Criterion validity

The extent to which the results of an experiment are correlated with those of other experiments already known to be valid. In the case of e.g. relevance assessments, if results from crowdsourced ground truth turn out to correlate well with results from earlier expert-established ground truth, the suitability of the corresponding crowdsourcing platform as a

scalable and less time-consuming ground truthing platform is strenghtened. An investigation like this has e.g. been done in [31] for the MIREX Audio Music Similarity and Retrieval task.

Internal validity

The extent to which the conclusions of an experiment can be rigorously drawn from the experimental design followed, and not from other factors unaccounted for. An optimal combination of musical attributes (e.g. good voice, catchy tune) will only partially explain high sales numbers for an artist; next to this, contextual aspects (such as recent high-profile appearances) will also play a role.

External validity

The extent to which the results of an experiment can be generalized to other populations and experimental settings. Of all the validity types mentioned here, issues with external validity may be the most concretely recognized in the MIR community at this moment. For example, many mid-level feature representations and assumptions in the MIR field have been modeled for Western popular music, but turn out not to be a good fit for other types of music: e.g. many classical music pieces do not have a constant tempo or steady beat, and an equal-tempered 12-tone chroma representation is not very well suited to capture the traditional music of other cultures.

Conclusion validity

The extent to which the conclusions drawn from the results of an experiment are justified. A notorious example is the claim that successful published work 'closed or bridged the *semantic gap*' (which will be discussed in more detail in the following section) — while indeed, low-level features often do not match high-level concepts, cases in which a better correspondence between these two levels is found frequently deal with domain-specific cases, and do not address any fundamental and generalized 'understanding' problems that a 'semantic gap' would imply. In addition, the whole metaphor of a semantic gap may not be appropriate; this will be addressed in the following section as well.

As we showed, meta-evaluation principles can readily be applied to many realistic MIR cases. By applying meta-evaluation principles, more insight can be gained into the scientific solidness of evaluation results, and because of this, the true intricacies of proposed systems will become clearer. This is very useful, since music data often is intangible data that is difficult to be understood, as we will discuss in the following section.

5 Opportunities for MIR: Universal Open Challenges

So far, we discussed transfer opportunities for two domains that are closely connected to the field of MIR. In this section, we will zoom out and take a higher-level perspective on open issues in the MIR field, and demonstrate that these are very similar to open fundamental issues as identified in the Content-Based Image Retrieval (CBIR) and Multimedia Information Retrieval (MMIR) communities, suggesting bridging opportunities for these fields and MIR.

5.1 The Nature of Music Data is Multifaceted and Intangible

Music is a peculiar data type. While it has communicative properties, it is not a natural language with referential semantics that indicate physically tangible objects in the world. One can argue that lyrics can contain such information, but these will not constitute music when considered in isolation.

The typical main representation of music is usually assumed to be audio or symbolic score notation. However, even such a representation in itself will not embody music as a whole, but rather should be considered a 'projection' of a musical object [75]. The composer Milton Babbitt proposed to categorize different music representations in three domains: 1) the *acoustic* or physical domain, (2) the *auditory* or perceived domain, and (3) the *graphemic* or notated domain. In [75], different transformations between these domains are mentioned: for example, a *transcription* will transform a mental image of music in the auditory domain to a notated representation in the graphemic domain, while a *performance* will transform the same mental image into an acoustic domain representation. The interplay between the three domains, in the presence of a human spectator, will establish experiences of the musical object, but that musical object itself remains an intangible, abstract concept.

Due to the multifaceted nature of music, and the strong dependence of experiences of music on largely black-boxed processes in the human auditory domain with strongly affective reactions, it is a very hard data type to grasp from a fundamental point of view. In an increasing amount of Music-IR tasks, we are typically not interested in precise (symbolic or digital) music encoding, nor in its sound wave dispersion behavior, but exactly in this difficult area of the effect music has on human beings, or the way humans interact with music. This poses challenges to the evaluation of automated methods: 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.

Issues with nonexistent ground truth, multifaceted representations and subjective and affective human responses are not new at all. In fact, they have been frequently mentioned in the CBIR and MMIR communities — although no clear and satisfying solution to them has been found yet.

5.2 Open Challenges are Shared Across Domains

In 2000 (incidentally, the year in which the first ISMIR conference was held), a seminal review [65] on content-based image retrieval (CBIR) was published, touching upon the state-of-the-art and outlining future directions. In this review, several trends and open issues were mentioned by the authors. It is striking to see how natural the following phrases read if transferred from the image to music processing domain, substituting 'CBIR' with 'MIR' and 'computer vision' with 'signal processing':

- The wide availability of digital sensors, the Internet, and the falling price of storage devices were considered as the *driving forces* for rapid developments in CBIR. However, more precise foundations would be desired, indicating what problem exactly is to be solved, and whether proposed methods would perform better than alternatives. A call was made for classification of usage-types, aims and purposes for the man-machine interface, domain knowledge, and database technology alike.
- *The heritage of computer vision*, from which CBIR developed, was considered to be an obstacle. CBIR is stronger about solving a general 'image understanding' problem and

evaluating results in terms of a user-defined ground truth than about providing algorithms with 100% segmentation accuracy according to a fully objective measure, which would be more typical of fundamental computer vision. Thus, in certain cases, goals could not exactly be taken over between these two related domains.

- Different goals and requirements in CBIR actually had *influence on computer vision* and (re)kindled interest in larger, dedicated datasets, weak segmentation and saliency, color image processing, and attention for invariance.
- It was argued that the notion of *similarity* should be considered from a human perspective. In addition, *learning* would be necessary to extend knowledge from partially labeled data to larger datasets.
- *Interaction* was mentioned as a major difference between CBIR and computer vision. Interaction and feedback mechanisms have been explored for a longer time in IR, but there are some fundamental differences between the two retrieval areas, especially in terms of query vs. result modalities. Visualization, so a move towards multimodal interfaces, was suggested as an important means to deal with this.
- Larger amounts of data increase the need for solid underlying *database* technology. Database research and CBIR traditionally have been separate fields, but were suggested to work together in this.
- *Evaluation* is a major issue. Results can be biased towards dataset composition, and it is hard to assess the 'difficulty' of a dataset. A call was made for reference standards such as TREC in Text IR. Furthermore, it was suggested to borrow concepts from the fields of psychological and social sciences.
- This review became particularly famous for coining the term *semantic gap* to indicate the mismatch between signal representations and analyses and the human assessments of their success. The authors wrote about resolving the gap by including *additional sources* of information. Here, insights from natural language processing and computer vision could be beneficial.

Many of these points still have largely remained unsolved. Eight years later, a survey in [12] still mentions user-focused (benchmark) evaluation as a future design goal, and application-oriented, domain-specific solutions as necessary ways to go in order to serve real-world needs.

With an increased interest in video data and multimodal approaches, part of the CBIR field merged into the MMIR field, where once again similar fundamental questions are mentioned. In [32], human-centered methods, multimedia-supported user-to-user collaboration, interactive search and agent interfaces, neuroscience and new learning models and folksonomies are pointed out as open future directions to study. The 'Holy Grail of Multimedia Information Retrieval', *getting the access to the content we like quickly and easily whenever we like it and wherever we are* [22], has not been found yet.

It is very striking to consider the open challenges mentioned above alongside the open challenges as identified at the occasion of the 10th anniversary of the ISMIR conference:

- 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.

In all cases, we identify a need for increased user involvement and interaction, understanding of the data while avoiding dataset bias, and the inclusion of multiple available information sources as main open challenges to pay attention to. Actually, even in the well-established IR field, involving the user is no common practice yet [3].

In both MMIR and MIR, it already has been hypothesized [49, 75] that a true semantic gap that can be 'crossed' through rigid algorithmic approaches is an unrealistic metaphor, and that human and cognitive approaches are necessary in any solution that is to be successful. The intangible and abstract nature of music data has strong potential to urgently push research into user-centered and multimodal approaches going towards this direction [33]. Thus, also in this area, we see opportunities, and even a potential flagship role, for MIR work to become of inspirational value to work in neighboring domains.

6 Conclusions

In this chapter, we discussed several methodology transfer opportunities for MIR. We first gave examples of MIR analogues to existing ASR tasks and discussed how MIR findings have benefited ASR the other way around. Subsequently, we mentioned current and promising influences from IR to MIR. Finally, we compared fundamental open challenges within MIR to those that have been mentioned, but never satisfyingly solved yet, in the CBIR and MMIR fields. Here, we argued that music data can be the key to finally address these challenges.

It is our intention that this chapter can serve as an inspirational guide, especially to researchers that are situated on the interfaces between different domains. We hope that increased bridge-building and knowledge exchanging between the domains will be capable of pushing research within these domains beyond limits and boundaries encountered so far.

References

1 E. Allamanche, J. Herre, B. Fröba, and M. Cremer. AudioID: Towards Content-Based Identification of Audio Material. In *Proc. 110th AES Convention, Amsterdam, NL*, 2001.

2 M. A. Bartsch and G. H. Wakefield. Audio Thumbnailing of Popular Music Using Chroma-based Representations. *IEEE Trans. on Multimedia*, 7(1):96–104, Feb. 2005.

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

4 W. Campbell, D. E. Sturim, D. Reynolds, and A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *Proc. of ICASSP*, pages 97–100, 2006.

5 C. Cao and M. Li. ThinkIT Submissions for MIREX2009 Audio Music Classification and Similarity Tasks. In *Music Information Retrieval Evaluation eXchange (MIREX)*, 2009.

6 C. Charbuillet, D. Tardieu, and G. Peeters. GMM Supervector for Content Based Music Similarity. In *Proc. of DAFx*, pages 1–4, 2011.

7 Z.-S. Chen, J.-S. Jang, and C.-H. Lee. A kernel framework for content-based artist recommendation system in music. *IEEE Transactions on Multimedia*, 2011. to appear.

8 H.-T. Cheng, Y.-H. Yang, Y.-C. Lin, I.-B. Liao, and H. Chen. Automatic chord recognition for music classification and retrieval. In *Proc. of ICME*, pages 1505–1508, 2008.

9 H. Christensen, J. Barker, N. Ma, and P. Green. The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments. In *Proc. of Interspeech*, pages 1918–1921, Makuhari, Japan, 2010.

10 R. Daido, S.-J. Hahm, M. Ito, S. Makino, and A. Ito. A System for Evaluating Singing Enthusiasm for Karaoke. In *Proc. of ISMIR*, pages 31–36, Miami, FL, USA, 2011.

11 D. Damm, H. Grohganz, F. Kurth, S. Ewert, and M. Clausen. SyncTS: Automatic synchronization of speech and text documents. In *Proceedings of the AES 42nd International Conference Semantic Audio*, Ilmenau, Germany, 2011.

12 R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys*, 40(2), 2008.

13 J. S. Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoust. Sci. & Tech.*, 29(4):247–255, 2008.

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

15 S. Ewert and M. Müller. Score-Informed Voice Separation For Piano Recordings. In *Proc. of ISMIR*, pages 245–250, Miami, FL, USA, 2011.

16 C. Fremerey, M. Müller, F. Kurth, and M. Clausen. Automatic mapping of scanned sheet music to audio recordings. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, September 2008.

17 H. Fujihara, M. Goto, and J. Ogata. Linking Lyrics: A Method for Creating Hyperlinks Between Phrases in Song Lyrics. In *Proc. of ISMIR*, pages 281–286, 2008.

18 J. Gemmeke, T. Virtanen, and A. Hurmalainen. Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2067–2080, 2011.

19 J. F. Gemmeke, T. Virtanen, and A. Hurmalainen. Exemplar-Based Speech Enhancement and its Application to Noise-Robust Automatic Speech Recognition. In *Proc. of CHiME Workshop*, pages 53–57, Florence, Italy, 2011.

20 M. Grimm and K. Kroschel. Evaluation of natural emotions using self assessment manikins. In *Proc. of ASRU*, pages 381–385. IEEE, 2005.

21 J. Han and C.-W. Chen. Improving melody extraction using probabilistic latent component analysis. In *Proc. of ICASSP*, pages 33–36, 2011.

22 A. Hanjalic, R. Lienhart, W.-Y. Ma, and J. R. Smith. The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away? *Proc. IEEE*, 96(4):541–547, April 2008.

23 A. Haubold and J. R. Kender. Alignment of speech to highly imperfect text transcriptions. In *ICME*, pages 224–227, 2007.

24 R. Hennequin, B. David, and R. Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *Proc. of ICASSP*, pages 45–48, Prague, Czech Republic, 2011.

25 N. Hu, R. B. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *in Proc. IEEE WASPAA*, pages 185–188, 2003.

26 X. Hu and J. S. Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proc. Joint Conference on Digital Libraries (JCDL)*, pages 159–168, Gold Coast, Queensland, Australia, 2010.

27 X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann. The 2007 MIREX Audio Mood Classification Task: Lessons Learned. In *Proc. of ISMIR*, pages 462–467, Philadelphia, USA, 2008.

28 J. Keshet, D. Grangier, and S. Bengio. Discriminative keyword spotting. *Speech Communication*, 51:317–329, 2009.

29 M. Khadkevich and M. Omologo. Use of hidden Markov models and factored language models for automatic chord recognition. In *Proc. of ISMIR*, pages 561–566, 2009.

30 T. Langlois and G. Marques. Automatic Music Genre Classification Using a Hierarchical Clustering and a Language Model Approach. In *Proc. of First International Conference on Advances in Multimedia*, pages 188–193, 2009.

31 J. H. Lee. Crowdsourcing Music Similarity Judgments using Mechanical Turk. In *Proc. ISMIR*, pages 183–188, August 2010.

32 M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-Based Multimedia Information Retrieval: State of the Art and Challenges. *ACM Trans. Multimedia Computing, Communications and Applications*, 2(1):1–19, 2006.

33 C. C. S. Liem, M. Müller, D. Eck, G. Tzanetakis, and A. Hanjalic. The Need for Music Information Retrieval with User-Centered and Multimodal Strategies. In *Proc. 1st Int. ACM workshop on MIR with User-Centered and Multimodal Strategies (MIRUM)*, pages 1–6, November 2011.

34 B. Logan. Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*, 2000.

35 B. Ma and H. Li. Text-Independent Speaker Recognition. In H. Li, K.-A. Toh, and L. Li, editors, *Advanced Topics in Biometrics*. World Scientific Publishing Co., 2011.

36 B. McFee and G. Lanckriet. The Natural Language of Playlists. In *Proc. of ISMIR*, pages 537–542, Miami, FL, USA, 2011.

37 A. Mesaros and T. Virtanen. Automatic alignment of music audio and lyrics. In *DAFX08*, 2008.

38 A. Mesaros and T. Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009. Article ID 546047.

39 M. Mohri, P. Moreno, and E. Weinstein. Robust Music Identification, Detection, and Analysis. In *Proc. of ISMIR*, Vienna, Austria, 2007.

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

41 M. Müller and D. Appelt. Path-constrained partial music synchronization. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, 2008.

42 M. Müller and F. Kurth. Towards Structural Analysis of Audio Recordings in the Presence of Musical Variations. *EURASIP Journal on Applied Signal Processing*, 2007(Article ID 89686):18 pages, January 2007.

43 M. Müller, F. Kurth, and M. Clausen. Audio matching via chroma-based statistical features. In *Proc. ISMIR, London, GB*, 2005.

44 M. Müller, H. Mattes, and F. Kurth. An Efficient Multiscale Approach to Audio Synchronization, 2006.

45 G. J. Mysore and P. Smaragdis. A Non-Negative Approach to Semi-Supervised Separation of Speech from Noise with the Use of Temporal Dynamics. In *Proc. of ICASSP*, pages 17–20, Prague, Czech Republic, 2011.

46 J. Nam, J. Ngiam, H. Lee, and M. Slaney. A Classification-Based Polyphonic Piano Transcription Approach Using Learned Feature Representations. In *Proc. of ISMIR*, pages 175–180, Miami, FL, USA, 2011.

47 M. Ogihara and T. Li. N-gram chord profiles for composer style representation. In *Proc. of ISMIR*, pages 671–676, 2008.

48 A. Ozerov, C. Févotte, and M. Charbit. Factorial scaled hidden Markov model for polyphonic audio representation and source separation. In *Proc. of WASPAA*, pages 121–124, Mohonk, NY, United States, 2009.

49 T. 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.

50 G. Poliner and D. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

51 L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, united states ed edition, Apr. 1993.

52 L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

53 B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh. Non-negative matrix factorization based compensation of music for automatic speech recognition. In *Proc. of Interspeech*, pages 717–720, Makuhari, Japan, 2010.

54 D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 3(1):72–83, January 1995.

55 J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

56 R. Scholz, E. Vincent, and F. Bimbot. Robust modeling of musical chord sequences using probabilistic N-grams. In *Proc. of ICASSP*, pages 53–56, 2009.

57 B. Schuller, C. Hage, D. Schuller, and G. Rigoll. "Mister D.J., Cheer Me Up!": Musical and Textual Features for Automatic Mood Classification. *Journal of New Music Research*, 39(1):13–34, 2010.

58 B. Schuller, B. Hörnler, D. Arsić, and G. Rigoll. Audio Chord Labeling by Musiological Modeling and Beat-Synchronization. In *Proc. of ICME*, pages 526–529, New York, NY, July 2009. IEEE, IEEE.

59 B. Schuller, C. Kozielski, F. Weninger, F. Eyben, and G. Rigoll. Vocalist Gender Recognition in Recorded Popular Music. In *Proc. of ISMIR*, pages 613–618, Utrecht, The Netherlands, October 2010. ISMIR, ISMIR.

60 B. Schuller, F. Weninger, and J. Dorfner. Multi-Modal Non-Prototypical Music Mood Analysis in Continuous Space: Reliability and Performances. In *Proc. of ISMIR*, pages 759–764, Miami, FL, USA, 2011.

61 J. Serrà. A Qualitative Assessment of Measures for the Evaluation of a Cover Song Identification System. In *Proc. ISMIR*, pages 319–322, September 2007.

62 J. Serrà, E. Gómez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Trans. on Audio Speech and Language Processing*, 16(6):1138–1152, August 2008.

63 M. Slaney. Precision-Recall is Wrong for Multimedia. *IEEE Multimedia*, 18(3):4–7, 2011.

64 P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 177–180, New Paltz, NY, USA, 2003.

65 A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. 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.

66 S. Steidl, B. Schuller, A. Batliner, and D. Seppi. The Hinterland of Emotions: Facing the Open-Microphone Challenge. In *Proc. of ACII*, pages 690–697, Amsterdam, The Netherlands, 2009.

67 J. Urbano. Information Retrieval Meta-Evaluation: Challenges and Opportunities in the Music Domain. In *Proc. ISMIR*, pages 609–614, October 2011.

68 E. Vincent, N. Bertin, and R. Badeau. Harmonic and Inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch Transcription. In *Proc. of ICASSP*, pages 109–112, 2008.

69 T. Virtanen. Speech Recognition Using Factorial Hidden Markov Models for Separation in the Feature Space. In *Proc. of INTERSPEECH*, pages 1–4, Pittsburgh, PA, USA, 2006.

70 D. von Zeddelmann, F. Kurth, and M. Müller. Perceptual Audio Features for Unsupervised Key-Phrase Detection. In *Proc. IEEE ICASSP*, Dallas, TX, USA, Mar. 2010.

71 A. Wang. An Industrial Strength Audio Search Algorithm. In *International Conference on Music Information Retrieval, Baltimore*, 2003.

72 B. Weiss and F. Burkhardt. Voice attributes affecting likability perception. In *Proc. of INTERSPEECH*, pages 2014–2017, 2010.

73 F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll. The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments. In *Proc. of CHiME Workshop*, pages 24–29, Florence, Italy, 2011.

74 F. Weninger, M. Wöllmer, and B. Schuller. Automatic Assessment of Singer Traits in Popular Music: Gender, Age, Height and Race. In *Proc. of ISMIR*, pages 37–42, Miami, FL, USA, 2011.

75 G. A. Wiggins, D. Müllensiefen, and M. T. Pearce. On the non-existence of Music: Why Music Theory is a figment of the imagination. *Musicae Scientiae*, Discussion Forum 5:231–255, 2010.

76 J. Wilpon, L. Rabiner, C.-H. Lee, and E. Goldman. Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(11):1870–1878, 1990.

77 M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. Interspeech*, pages 597–600, Brisbane, 2008.

78 M. Wöllmer, F. Weninger, F. Eyben, and B. Schuller. Acoustic-Linguistic Recognition of Interest in Speech with Bottleneck-BLSTM Nets. In *Proc. of INTERSPEECH*, pages 77–80, Florence, Italy, 2011.

79 Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. Chen. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):448–457, 2008.

80 K. Yoshii and M. Goto. A Vocabulary-Free Infinity-Gram Model for Nonparametric Bayesian Chord Progression Analysis. In *Proc. of ISMIR*, pages 645–650, Miami, FL, USA, 2011.

81 C. H. You, K.-A. Lee, and H. Li. An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition. *IEEE Signal Processing Letters*, 16(1):49–52, 2009.

82 S. J. Young, G. Evermann, M. J. F. Gales, D. Kershaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK book version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006.

Grand Challenges in Music Information Research

Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
m.goto@aist.go.jp

Abstract

This paper discusses some grand challenges in which music information research will impact our daily lives and our society in the future. Here, some fundamental questions are how to provide the best music for each person, how to predict music trends, how to enrich human-music relationships, how to evolve new music, and how to address environmental, energy issues by using music technologies. Our goal is to increase both attractiveness and social impacts of music information research in the future through such discussions and developments.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems, I.5 Pattern Recognition

Keywords and phrases Music information research, grand challenges, music processing, music signals

Digital Object Identifier 10.4230/DFU.Vol3.11041.217

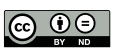
1 Introduction

Music information research is gaining a lot of attention [15, 7, 11, 2]. It has a long history as shown by attempts to use a computer to compose music from the time of the invention of the computer, such as the "Illiac Suite for String Quartet" of 1957. Results from music information research have spread widely throughout society, including synthesizers which have become essential for the production of popular music, and music distribution services over mobile phones. The field of music information research covers all aspects of music and all aspects of people's music activities, and is related to a variety of topics such as signal processing, transcription, sound source segregation, identification, analysis, understanding, retrieval, recommendation, classification, distribution, synchronization, conversion, processing, summarization, composition, arrangement, songwriting, performance, accompaniment, score recognition, sound synthesis, singing synthesis, generation, assistance, encoding, visualization, interaction, user interfaces, databases, annotation, and social tags related to music. The aims of music information research as an academic field are to study mechanisms for

- listening to and understanding music,
- creating and performing music,
- distributing, retrieving, and recommending music,
- communication between people through music, and
- qualities intrinsic to music

from the viewpoints of science (revealing the truth) and engineering (making useful systems).

The importance of music information research was not recognized until the 1990s, however. This was transformed dramatically after 2000 when the general public started listening to music on computers in daily life. It is now widely known as an important research field, and new researchers are continually joining the field worldwide. Although music information research sometimes needed an argument to be recognized as serious research instead of



Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 217–226

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

research for fun more than a decade ago, such misconceptions become a thing of the past. This change has been caused by the fact that the general public is aware that all music will eventually be digitized, created, distributed, used, shared, etc. There will be further demand for new music listening interfaces, retrieval, and recommendations. Academically, one of the reasons many researchers are involved in this field is that the essential unresolved issue is the understanding of complex musical audio signals that convey content by forming a temporal structure while multiple sounds are interrelated [11, 3, 4, 15]. Additionally, there are still appealing unresolved issues that have not been touched yet, and the field is a treasure trove of research themes.

This paper discusses some grand challenges that could further increase both the attraction and social impacts of music information research in the future. Please note that some discussions in this paper are intentionally provocative to trigger controversial discussions and stimulate new ideas.

2 Grand Challenges

How can music information research contribute to building a better world and making people happy? How can it contribute to solving the global problems our worldwide society faces? This paper discusses some grand challenges that could be tackled by music information research and could also convince the general public that this research has social impacts for a better, sustainable world and is really important for enriching their lives.

2.1 Can music information research provide music or music videos optimized to the individual?

The goal here is to provide the best music for each person by generating or finding appropriate context-aware music. Music preferences vary from person to person, and even the same person may want to listen to different music (or watch music videos) depending on their situation or mood. If it is technically possible to automatically generate (compose) optimal songs or select such songs from a huge lineup of existing music according to such preferences, situations, or moods, people could not stop using this technology that always provides super happiness and joy. This would have a big impact on society, though such a technology would be controversial if people are really addicted to it. In order to achieve this, technology that is able to understand music and music videos in the same way people do is important. Current technology is not able to do this in terms of

- the ability of understanding people's preferences and situations,
- the quality of automatically generated music or music videos,
- the accuracy of music selection (retrieval and recommendation), or
- the depth of automatic music understanding.

However, there is room for discussion regarding whether a completely automated system is the best. For example, an approach of making an interactive system that assists people's activities is also appealing [6].

Because it is difficult to automatically generate new music or music videos from scratch, the music or video provided could be *2nd generation (secondary or derivative) content*. In the 2nd generation content, musical elements and ideas of existing music or video called *1st generation (primary or original) content* [16] are reused in the creation of new songs. Even 3rd, 4th, or *N-th generation content* can be considered by reusing the generated content again and again as shown in Figure 1. Satoshi Hamano named this style of content creation *N-th order derivative creation*. The reuse or customization of existing music might be a more

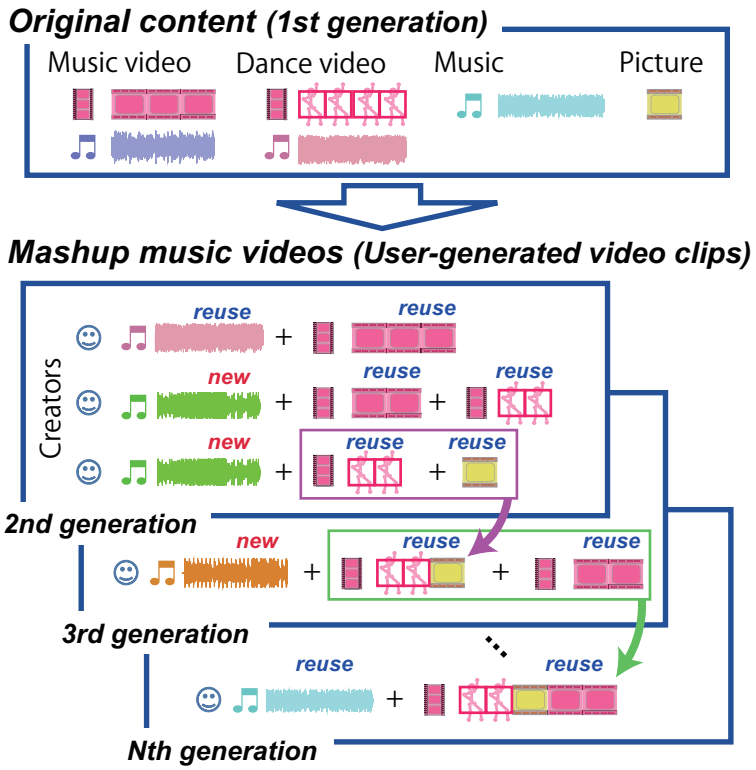


Figure 1 Generation of mashup music videos (user-generated music video clips) by reusing existing original content [16].

natural approach for the future. For example, "mash-ups" that ingeniously combine and mix different songs, and "touch-ups" (customizations) that modify or customize elements of existing songs (changing the timbre, phrase, and volume balance of singing voices and musical instruments) [6] are important in the discussion of music creation. In fact, in recent years, there have been increasing activities to intentionally provide songs and elements so that other people can use them for the N-th order derivative creation [14, 8].

2.2 Can music trends be predicted?

The goal here is to predict music trends by predicting hit songs to cause or prevent a "music pandemic". Is it technically possible to predict hit songs? Alternatively, is it technically possible to provide reasons why a song is not selling? There are actual studies on "hit song science" [17], but technology that is able to predict global or local trends with a high level of precision has not yet been achieved. Prediction of trends is difficult to derive from only the content of music, and it is necessary to globally and exhaustively incorporate information on the Web as social information in order to achieve results that could not be achieved using only technology for analyzing and understanding audio signals [19].

Putting aside the pros and cons of the surveillance society aspect, such trend prediction would become more feasible if it were possible to obtain a worldwide history of what kinds of music everyone is listening to. That is, through the further spread of music distribution technology, it will become possible to record the history of all music playback and sharing this while maintaining anonymity. By making it possible to record the history of what individuals

listen to using automatic song identification technology even in live performances [20, 12, 18], it is likely that it will be possible to predict music trends with a high level of precision. However, at the same time, it is interesting to speculate whether, once it becomes possible to provide music optimized for individuals, diversification of music will accelerate and trends will become less likely to occur (thus preventing a "music pandemic"), or people will want to hear what others are listening to, resulting in huge trends (as if a "music pandemic" had occurred).

2.3 Can the relationship between people and music be made richer?

The goal here is to enrich human-music relationships by reconsidering the concept of originality. Digitizing all music from past to present will enable humankind to instantly access all music for the first time ever. Moreover, music will continue to accumulate. The number of accessible songs has monotonically increased, and the number of musical pieces registered on flat-rate music distribution services such as Napster 2.0 has reached 15 million musical pieces. Access will become even easier in the future with progress in music information retrieval and recommendation technology. This itself is historically inevitable, and is desirable as it will make people's music lifestyles more convenient. However, music information research holds the key to whether this will eventually enrich the relationship between people and music.

In the past, new artists needed to try to ensure that their songs were not buried among all the other songs that were on the market, but in the future, it may become even more difficult to get people to listen to music because it is buried among an enormous number of all songs from the past to the present. Moreover, once it becomes possible to automatically compute similarities with all past songs in terms of partial elements such as melody, lyrics, chord progression and arrangement, it will become clear that all songs may have similar aspects to other songs. This is because all creations are affected by other works on a subconscious level. In some cases, it may be technically possible to point out past songs that are partially similar to a song that has just been created. It will be interesting to see how this transforms copyright concepts, and the concept of the originality of music may need to be reconsidered.

So does this mean that human will be unable to overcome the music of the past and lose the will to create new music? Will new music no longer be needed? I don't think so. Essentially, the important things about music are not originality and copyright, but rather how it inspires and makes people happy, and its overall appeal and quality as a work of art. Furthermore, the joy of expression itself is another driving force behind music creation. We may see the arrival of an era in which we go back to the origin of music in a time when it could only be enjoyed in a live concert without the ability to record as more emphasis is placed on using music to bring joy and pleasure to people "here and now." Technological advances could bring about a new music culture that is more centered on emotional, touching experiences.

2.4 Will music information research bring about the evolution of music itself?

The goal here is to push new music evolution forward by enabling new music representations to emerge or enhancing human abilities of enjoying music. The emergence of new technology has already created new musical expressions. This will inevitably continue to create new musical expressions in the future. For example, automatic pitch-correction technology of vocals is already being used on a routine basis in the production of commercial music (popular music, in particular). It has become an absolute necessity for correcting pitch at points in a

Figure 2 Singing synthesis software *Hatsune Miku* with a cute synthesized voice and an illustration of a cartoon girl. (Courtesy of © CRYPTON FUTURE MEDIA, INC.)

song where the singer is less than skillful and for making corrections to achieve a desired effect. Furthermore, since 2007, singing synthesis technology represented by Yamaha's VOCALOID [10] has gained much attention in Japan. Both amateur and professional musicians have started to use singing synthesizers as their main vocals, and songs sung by computer singers rather than human singers have become popular and are now being posted in large numbers on video sharing services likes *Nico Nico Douga* (`http://www.nicovideo.jp/video_top/`) in Japan and *YouTube* (`http://www.youtube.com`). Even compact discs featuring compilations of songs created using singing-synthesis technology are often sold and appear on popular music charts in Japan [9]. In particular, *Hatsune Miku* [14, 1] is the name of the most popular software package based on VOCALOID and has a cute synthesized voice with an illustration of a cartoon girl as shown in Figure 2. Although Hatsune Miku is a virtual singer, she has already had live concerts with human musicians in Japan, USA, and Singapore (Figure 3). As music synthesizers generating various instrumental sounds are already widely used and have become indispensable to popular music production, it is historically inevitable that singing synthesizers will become more widely used and likewise indispensable to music production. Initially, synthesizers could easily be distinguished from the sound of natural instruments, and this itself led to the creation of unique expressions, but now the quality is high enough that they cannot be differentiated by the general public, and they are used in the majority of popular music. There is no reason that the same will not happen for singing synthesis. The only uncertainty is how soon this will happen.

I hypothesize that the complexity of music created by humankind as audio signals is monotonically increasing. However, there is a limit to the complexity the general public finds enjoyable, and increases in complexity using the approaches of contemporary music have had difficulty in gaining popularity. I believe that the "mash-ups" mentioned earlier hold one of the keys to the next evolution of music from this perspective. Mash-ups are a music production technique in which multiple songs (or their components such as only

Figure 3 Live concert by Hatsune Miku at MIKUNOPOLIS 2011 [13] in Los Angeles, USA on July 2nd, 2011. (Courtesy of © CRYPTON FUTURE MEDIA, INC. and © MIKUNOPOLIS 2011)

the vocals or accompaniment) are used as material to be mixed together and combined as if they were parts of the same song from the beginning. By referring to the musical memory already in the mind of the listener, these mash-ups are able to raise the level of complexity acceptable for enjoyment while retaining popularity. In the days when there were no electronic instruments, it was only possible to create music based on units of single notes (individual instrumental notes) on a musical score, but advances in technology have made it possible to produce music using musical fragments of several bars (one phrase) as units or loop material. Mash-ups are musical productions using whole songs as units or material, making it easier to achieve complex audio signals that would be inconceivable when creating a song from scratch. From the viewpoint of listeners, on the other hand, when the songs used as material to be mixed together are already in the memory, they can enjoy songs that would normally be too complex to enjoy.

Has the tempo of music also monotonically increased throughout the history of humankind? If that is the case, the same song would be shorter in length if the tempo were increased, and the number of songs that can be listened to per unit of time can be expected to increase. This is convenient for the "era of access to an enormous number of songs" mentioned above. If that is the case, how fast can songs be made to still be enjoyed by the human brain? Furthermore, what kinds of technologies can be used to assist and train this? It is intriguing whether the human hearing and capability of the brain are able to keep pace and improve when the tempo is systematically increased by 5 BPM[1] every year by (worldwide) laws. I know this idea is especially provocative, but it is worth thinking about the evolution of music in a think-outside-the-box way.

[1] BPM (Beats Per Minute) is a unit indicating the tempo of a performance based on the number of beats in a minute.

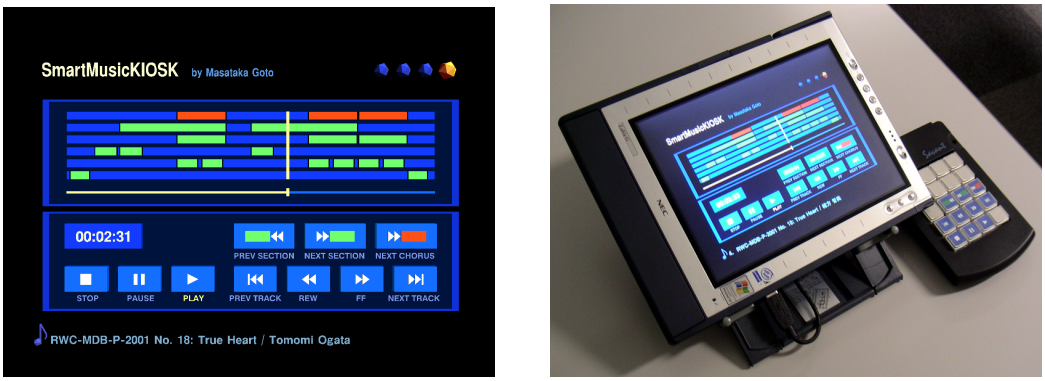


Figure 4 SmartMusicKIOSK screen display. This ia a music listening station with a chorus-search function. The lower window presents the playback operation buttons and the upper window provides a visual representation of a song's contents. A user can actively listen to various parts of a song while moving back and forth as desired on the visualized song structure (the "music map" in the upper window).

2.5 Can music information research contribute to addressing environmental issues and energy issues?

The goal here is to contribute to solving the global problems our worldwide society faces. Environmental issues can be addressed by contributing to a reduction in the use of resources through efforts to increase online music that eliminates the need for physical media (tapes, records, CDs and DVDs). Advances in technology have brought us to an era in which "music" as packaged media can be seen as "information" not affected by physical media, but physical media are still being distributed. Just as overwhelming convenience brought about the transition from record distribution to CD distribution, overwhelming convenience is required for the transition from distribution of physical media using many environmental resources to the distribution of information. Music understanding technology is one means of providing this convenience, and convenience is expected to be improved in various ways such as in "Active Music-Listening Interfaces" [6]. For example, SmartMusicKIOSK [5], an Active Music-Listening Interface with an automatic chorus-section detection technology enables automatic visualization of song structure to listen to parts that are interesting (Figure 4).

With regard to the energy issue, music can be considered a form of high-quality entertainment that does not require much energy. The resources and energy required for music production is less than for production of motion pictures, and will further decrease significantly through the spread of digital music production environments. The *N-th order derivative creation* and mash-ups that reuse (or "recycle") existing songs are also positioned as energy-efficient ways to produce music, and the development of technologies to assist such production is vital. Furthermore, music can be listened to repeatedly, and it is possible to listen to the same song many times. In fact, repeated listening essentially enables the listener to notice a song's appeal. Listening support such as the "Active Music-Listening Interfaces" [6] mentioned above, which enable a deeper understanding of existing music, can contribute to this. Furthermore, advances in music distribution technology will lower distribution costs, and if it is possible to listen to only one's preferred music thanks to advances in music information retrieval and recommendation technology, energy spent on music one is not interested can be reduced. This could be named as "energy-conscious" music production and appreciation. "Music happiness per energy" can thus be increased.

3 Conclusion

In the future, what will be necessary to further increase the appeal of music information research in addition to addressing the above grand challenges?

First, we must develop technology that contributes to building a better world and making people happy, and is essential for society. We hold the key to the creation of a mentally rich future society, and it is vital that academia and industry are seriously engaged in interaction and mutual development. We would need to further discuss what should be done to contribute to the advancement of the music industry and the creation of new industries, and how a contribution can be made to the future of music production and music appreciation.

Second, the importance of our research field must be emphasized as must the further understanding that additional investment in research and development is required. To do this, we must produce researchers who will generate a variety of appealing research results of the highest quality, and also make an effort to talk about our dreams for the future. Such activities will lead to large projects and a diversity of funding, and it would be good to promote great advances in research with a variety of financial backing.

Third, we must promote the field of music information research much more, and make it easy for anyone to feel comfortable participating in it. I would like to expand the research field as a whole, to make possible exciting results from a more diverse range of research.

This paper was written with the aim of contributing the three points above, and I hope that such discussions will continue to be active throughout the field as a whole. However, this must not involve moving in the direction of creating the shell around "music information research" and becoming stuck inside it. What is necessary is a range of activities that span boundaries between fields and reorganize learning from a broader perspective. The field of music information research is also expected to make great strides such as merging with spoken language processing and image processing. I look forward to what the future holds in ten years time.

References

1 Cabinet Office, Government of Japan. Virtual idol. In *Highlighting JAPAN through images*, volume 2, pages 24–25, 2009.
`http://www.gov-online.go.jp/pdf/hlj_img/vol_0020et/24-25.pdf`

2 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, 2008.

3 Masataka Goto. Music scene description project: Toward audio-based real-time music understanding. In *Proc. of the 3rd International Conference on Music Information Retrieval (ISMIR 2003)*, pages 231–232, 2003.

4 Masataka Goto. A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329, 2004.

5 Masataka Goto. A chorus-section detection method for musical audio signals and its application to a music listening station. *IEEE Trans. on ASLP*, 14(5):1783–1794, 2006.

6 Masataka Goto. Active music listening interfaces based on signal processing. In *Proc. of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, pages 1441–1444, 2007.

7 Masataka Goto and Keiji Hirata. Invited review: Recent studies on music information processing. *Acoustical Science and Technology (edited by the Acoustical Society of Japan)*, 25(6):419–425, 2004.

8 Masahiro Hamasaki, Hideaki Takeda, and Takuichi Nishimura. Network analysis of massively collaborative creation of multimedia contents: Case study of Hatsune Miku videos on Nico Nico Douga. In *Proc. of the 1st international conference on Designing interactive user experiences for TV and video (uxTV' 08)*, pages 165–168, 2008.

9 Hideki Kenmochi. VOCALOID and Hatsune Miku phenomenon in Japan. In *Proc. of the First Interdisciplinary Workshop on Singing Voice (InterSinging 2010)*, pages 1–4, 2010.

10 Hideki Kenmochi and Hayato Ohshita. Vocaloid – commercial singing synthesizer based on sample concatenation. In *Proc. of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, pages 4011–4010, 2007.

11 Anssi Klapuri and Manuel Davy, editors. *Signal Processing Methods for Music Transcription*. Springer, 2006.

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

13 Crypton Future Media. Mikunopolis in Los Angeles. `http://mikunopolis.com`.

14 Crypton Future Media. What is the "HATSUNE MIKU movement"? `http://www.crypton.co.jp/download/pdf/info_miku_e.pdf`.

15 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.

16 Tomoyasu Nakano, Sora Murofushi, Masataka Goto, and Shigeo Morishima. DanceReProducer: An automatic mashup music video generation system by reusing dance video clips on the web. In *Proc. of the 8th Sound and Music Computing Conference (SMC 2011)*, pages 183–189, 2011.

17 Francois Pachet and Pierre Roy. Hit song science is not yet a science. In *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, pages 355–360, 2008.

18 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, October 2008.

19 Malcolm Slaney. Web-scale multimedia analysis: Does content matter? *IEEE MultiMedia*, 18(2):12–15, 2011.

20 Avery Wang. The Shazam music recognition service. *Communications of the ACM*, 49(8):44–48, 2006.

Music Information Technology and Professional Stakeholder Audiences: Mind the Adoption Gap

Cynthia C. S. Liem*[1], Andreas Rauber[2], Thomas Lidy[2,3], Richard Lewis[4], Christopher Raphael[5], Joshua D. Reiss[6], Tim Crawford[4], and Alan Hanjalic[1]

1 **Multimedia Information Retrieval Lab**
Delft University of Technology, The Netherlands
c.c.s.liem@tudelft.nl, a.hanjalic@tudelft.nl
2 **Information Management and Preservation Lab**
Vienna University of Technology, Austria
rauber@ifs.tuwien.ac.at, lidy@ifs.tuwien.ac.at
3 **Spectralmind GmbH, Austria**
4 **Department of Computing**
Goldsmiths, University of London, United Kingdom
richard.lewis@gold.ac.uk, t.crawford@gold.ac.uk
5 **School of Informatics**
Indiana University, Bloomington, USA
craphael@indiana.edu
6 **Centre for Digital Music**
Queen Mary, University of London, United Kingdom
josh.reiss@eecs.qmul.ac.uk

Abstract

The academic discipline focusing on the processing and organization of digital music information, commonly known as Music Information Retrieval (MIR), has multidisciplinary roots and interests. Thus, MIR 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 MIR 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 MIR 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 contribution, we discuss MIR 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 MIR academics, influencing their reception of new MIR 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 MIR.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, Performing arts, K.4.3 Organizational Impacts, H.3.7 Digital Libraries–User issues

Keywords and phrases music information retrieval, music computing, domain expertise, technology adoption, user needs, cross-disciplinary collaboration

Digital Object Identifier 10.4230/DFU.Vol3.11041.227

* The work of Cynthia Liem is supported in part by the Google European Doctoral Fellowship in Multimedia.

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 227–246

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

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* (MIR). 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 MIR actually does not actively deal with retrieval, the field has alternatively been called *Music Information Research*, retaining the same acronym.

The largest MIR success story so far may have been in audio fingerprinting (e.g. [27]), which is widely adopted in today's consumer devices[1]. Academic MIR 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 MIR 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 MIR 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 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 [24]), they will mainly concern strategies to optimize effective usage of the technology, giving the user a

[1] It is not uncommon for an enthusiastic MIR 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

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 MIR and related engineering disciplines; if it is, it is the realm of library science experts[4], not of engineers.

Of course, not every MIR research project has the urgence to immediately culminate into a monetized end-user system. Nonetheless, the MIR 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 MIR 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 contribution, 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, the reception of an intelligent audio mixing system is described. Section 3 will subsequently describe a case study on the *Music Plus One* musical accompaniment system, and discuss prevalent lines of thought in classical musicianship.

The subsequent sections will deal with broader cross-disciplinary adoption and collaboration issues. For quite some time, MIR 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 4 will elaborate on this, elucidating how current musicological interests are different from the common assumptions in MIR. Finally, a very different, but important category of professional users and collaboration partners is formed by stakeholders and representatives in the music industry. Section 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.*

Our contribution will be concluded with a discussion in Section 6, in which common adoption issues will be summarized and recommendations are given to overcome them.

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.

[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.

2.1 "Is this a joke?"

In the automatic mixing work of Enrique Perez Gonzalez and Joshua D. Reiss [19, 20, 21], 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 comments 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, his editorial for the *Sound on Sound* magazine of October 2008 [28], 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 [18] 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 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

[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.

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

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.

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*) [22, 23], 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 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.

[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.

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 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.*

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

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

'liberal arts': first the *trivium* consisting of grammar, logic and rhetoric, and afterwards the *quadrivium* consisting of geometry, arithmetic, astronomy — and music [14].

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 [7].

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 MIR. These will be discussed in the following section.

4 Musicology

Within the scientific MIR 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 MIR has to offer musicology, or how to attract musicologists to the field (e.g. [8, 11, 12]), and a musicologist at an MIR event can

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

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 MIR 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 MIR 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 MIR and musicology can truly get closer to each other.

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 [4] 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 [10] addressed more recreational information seeking, some of her conclusions are nevertheless relevant to scholars, particularly that advanced MIR techniques are not often developed beyond proof-of-concept into practical, usable tools.

Barthet and Dixon [2] 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 MIR 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* [25], as well as the current of criticism in the digital humanities.

[12] It should be noted that at least eighteen review articles in *JAMS* also mention computational subjects.
[13] Sonic Visualiser is a tool for interactive sound analysis providing a variety of visualizations, annotation, and plugin analytical procedures.

For example, Unsworth [26] 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 [15] 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.

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 [7], emphasizing subjectivity and cultural context, and refuting objective, universal, 'scientific' views on music.

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 [3] 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 (MIR) 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 [31]) 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 MIR and musicology leads to a situation in which a lot of work that MIR researchers enthuse over is meaningless to musicologists. Frequently, the MIR technologists tend to focus on what seem, from a musicologist's perspective, to be more low-level 'problems' rather than higher level 'questions'.

[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.

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 MIR 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 which MIR 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 categorising 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.

4.3 Outlook for Musicology

Since it seems that present-day musicology fundamentally has other interests than MIR researchers would initially assume, where does all this leave the advocates of MIR to musicology? One approach which is being taken is to introduce more musically sophisticated topics of research into the MIR agenda. Particularly, Wiering is encouraging investigation into the broad topic of musical meaning using MIR techniques [29] and has, together with Volk, also been responsible for encouraging those working in MIR to find out more about contemporary musicology [30], arguing that it is a "founding discipline" of MIR.

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.

[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.

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

As an example, [9] 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 MIR community already, e.g. in [1, 13, 17].

At a more global level, the interest of contemporary musicology in contexts around musical practice resonates very well with the current interest in MIR 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 MIR 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 MIR, but actually have not been fully solved yet.

5 Music Industry: Findings from the CHORUS+ Think-Tank

If the goal of an MIR 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 MIR research.

In this section, the findings of both the survey and Think-Tank roundtable discussions relevant to the topic of this contribution will be presented. The full report on the Think-Tank,

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

as well as a full list of its participants, is available online [16]. 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).

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 a.o. 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;
- 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."

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

- "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 MIR research, since they largely overlap with current academic research interests in MIR. 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'.

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 MIR research perhaps going in too complex directions regarding this?

Following the questions posted above, the role of the so-called *long tail* was discussed [5]. 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.

MIR 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 MIR 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 MIR 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.

5.3 Technological or Business Model Issues?

In some cases, research and development (R&D) in MIR 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

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

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 MIR 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. 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.

5.4 Outlook for Industry

Current business models and legal issues seem to consider existing MIR technology to be sufficient for monetization purposes, and thus make it very hard for new and innovative MIR technology to get adopted. Does this mean that current MIR 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 MIR 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 MIR 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 MIR 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.

As a good demonstrator of the potential impact and success of MIR 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.

6 Discussion

In this contribution, multiple difficulties were pointed out regarding the adoption of MIR 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 MIR 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

[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.

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 MIR 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 [6], 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 MIR 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 MIR 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 MIR tools. While the music industry stakeholders purely focusing on sales may not be interested in novel MIR 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 MIR research directions and linked data.

A striking feature of the MIR 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 MIR 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.

Our contribution was meant to increase awareness on this topic and to give a warning to the enthusiastic MIR researcher. As we demonstrated, several reception and adoption issues are of fundamental nature and may be very difficult to overcome.

On the other hand, our contribution was certainly not meant as a discouragement. There are many promising (and possibly unexpected) MIR 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.

References

1 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)*, May 2011.

2 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 2011)*, Miami, USA, 2011.

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

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

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

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

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

8 Nicholas Cook. The Compleat Musicologist. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, London, UK, 2005.

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

10 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 2003)*, Houston, Texas, 2003.

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

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

12 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 2009)*, Kobe, Japan, 2009.

13 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 2009)*, Kobe, Japan, 2009.

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

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

16 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.

17 Cynthia C. S. Liem and Alan Hanjalic. Expressive Timing from Cross-Performance and Audio-based Alignment Patterns: An Extended Case Study. In *Proc. of the 12th Conf. of the Int'l Society for Music Information Retrieval (ISMIR 2011)*, Miami, USA, 2011.

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

19 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-07)*, Bordeaux, France, September 2007.

20 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 2008.

21 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 2008.

22 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 2001.

23 Christopher Raphael. Music Plus One and Machine Learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, Haifa, Israel, June 2010.

24 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. Schloss Dagstuhl - Leibniz Center für Informatik GmbH, 2012.

25 Charles Percy Snow. *The Two Cultures*. Cambridge University Press, 1993.

26 John Unsworth. New Methods for Humanities Research. The 2005 Lyman Award Lecture, November 2005.

27 Avery Li-Chun Wang. An Industrial-Strength Audio Search Algorithm. In *Proc. of the 4th Int'l Conf. on Music Information Retrieval (ISMIR 2003)*, Baltimore, USA, October 2003.

28 Paul White. Automation For The People. *Sound on Sound*, October 2008.

29 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.

30 Frans Wiering and Anja Volk. Musicology. Tutorial slides, ISMIR 2011.

31 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.