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

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 FOLLOW-UPS 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 FOLLOW-UPS 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 FOLLOW-UPS 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