Report from Dagstuhl Seminar 11041

Multimodal Music Processing

Edited by

Meinard Müller[1], Masataka Goto[2], and Simon Dixon[3]

1 **Saarland University and MPI Informatik - Saarbrücken, DE,** meinard@mpi-inf.mpg.de
2 **AIST – Ibaraki, JP,** m.goto@aist.go.jp
3 **Queen Mary University of London, GB,** simon.dixon@eecs.qmul.ac.uk

Abstract

From January 23 to January 28, 2011, the Dagstuhl Seminar 11041 “Multimodal Music Processing” was held at Schloss Dagstuhl – Leibniz Center for Informatics. During the seminar, we discussed various aspects of the automated processing of music-related documents. These documents may describe a musical work in different ways comprising visual representations (e. g., sheet music), symbolic representations (e. g., MIDI, tablatures, chords), acoustic representations (CD recordings), audio-visual representations (videos), or text-based metadata. In this report, we give an overview of the main contributions and results of the seminar. We start with an executive summary, which describes the main topics, goals, and group activities. Then one finds a list of abstracts giving a more detailed overview of the participants’ contributions as well as of the ideas and results discussed in the group meetings and panels of our seminar.

Seminar 23.–28. January, 2011 – http://www.dagstuhl.de/11041
1998 ACM Subject Classification H.5.5 Sound and Music Computing, I.5.4 Applications—Waveform analysis, H.5.1 Multimedia Information Systems, J.5 Arts and Humanities
Keywords and phrases Music information retrieval, music processing, multimodality, audio, sheet music, content-based analysis, signal processing, user interaction
Digital Object Identifier 10.4230/DagRep.1.1.68
Edited in cooperation with Sebastian Ewert

1 Executive Summary

Meinard Müller
Masataka Goto
Simon Dixon

License 🅭 Creative Commons BY-NC-ND 3.0 Unported license
© Meinard Müller. Masataka Goto, Simon Dixon

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. Furthermore, music notation can be encoded in text-based formats such as MusicXML, or symbolic formats such as MIDI. Music annotations, metadata, and social tags are also widely available. Beside textual data, increasingly more types of music-related multimedia data such as audio, image or video data are widely available. For example, there are numerous MP3 and CD audio recordings, digitized images of scanned sheet music, and an increasing number of video clips of music performances. In this seminar we discussed and studied various aspects of the

Except where otherwise noted, content of this report is licensed under a Creative Commons BY-NC-ND 3.0 Unported license

automated processing of music-related documents that vary in their modalities and formats, i.e., text, symbolic data, audio, image and video. Among others, the topics dealt with aspects of data analysis, retrieval, navigation, classification, and generation of music documents while considering music-specific characteristics, exploiting multiple information sources, and accounting for user-specific needs.

In this executive summary, we give a brief overview of the main topics addressed in this seminar. We start by briefly describing the background of the seminar participants and the overall organization. We then give an overview of the presentations and the results from working groups and panels. Finally, we reflect on the most important aspects of this seminar and conclude with future implications.

Participants, Interaction, Activities

In our seminar, we had 35 participants, who came from various countries around the world including North America (7 participants), Japan (4 participants), New Zealand, Singapore, and Europe (Austria, France, Germany, Netherlands, Portugal, Spain, United Kingdom). Most of the participants came to Dagstuhl for the first time and expressed enthusiasm about the open and retreat-like atmosphere. Besides its international character, the seminar was also highly interdisciplinary. While most of the participating researchers are working in computer science and its neighboring fields, a large number of participants also have a strong background in music and musicology. This made the seminar very special in having not only interactive and provoking scientific discussions, but also numerous social activities including common music making. One particular highlight of such social activities was a three-hour spontaneous concert on Thursday evening, where various participants performed in changing ensembles a wide variety of music including popular music, jazz, and classical music.

Overall Organization and Schedule

Dagstuhl seminars are known for having a high degree of flexibility and interactivity, which allow participants to discuss ideas and to raise questions rather than to present research results. Following this tradition, we fixed the schedule during the seminar asking for spontaneous contributions with future-oriented content, thus avoiding a conference-like atmosphere, where the focus is on past research achievements. The first day was used to let people introduce themselves and express their expectations and wishes for the seminar. We then had a brainstorming session on central topics covering the participants' interests while discussing the overall schedule and format of our seminar. In particular, we identified a total of six topics for discussion. For four of these topics, we divided into four groups, each group discussing one of the topics in greater depth in parallel sessions on Tuesday. The results and conclusions of these group meetings were then presented to the plenum on Wednesday. For the remaining two topics, we decided on having panel-like discussions within the plenum with introductory stimulus talks (Thursday). Finally, group and panel discussions were interleaved with regular sessions that allowed participants to present their personal research to the plenum. This mixture of presentation elements gave all participants the opportunity for presenting their ideas to the plenum while avoiding a monotonous conference-like presentation format.

Main Topics and Results

We discussed various topics that address the challenges of organizing, understanding, and searching music-related information in a robust, efficient, and intelligent manner. Here, a particular focus was put on music-specific aspects, the fusion of multiple information sources, as well as the consideration of user-specific needs. After a joint brainstorming session, we agreed on discussing six central topics which fitted in the overall theme of the seminar and reflected the participants' interests. We now give a brief summary of these topics, which were discussed within four parallel group meetings and two panels. Then, we give an overview of additional contributions made by the participants in the regular sessions of the seminar.

1. The *"Model"* group discussed the issue of how signal models can be developed that exploit multimodal information. Here, one main goal was to review strategies for combining different sources of information to support music analysis. In particular, various early and late fusion approaches were identified and advantages and weaknesses of the respective approaches were discussed. Particular attention was paid to the aspect of data uncertainty and its propagation in the fusion processes.
2. The *"User"* group addressed the topic of user-aware music information retrieval. Here, a central question was how contextual information can be integrated into the retrieval process in order to account for short-term user interests and long-term user behavior. Additionally, it was discussed how search engines may yield satisfying results in terms of novelty, popularity, and serendipity of the retrieved items.
3. The *"Symbol"* group discussed the question of how to bridge the gap between visual, symbolic, and acoustic representations of music. Here, particular attention was given to the problem referred to as *Optical Music Recognition* (OMR) with the goal of converting an image-based sheet music representation into a symbolic music representation where note events are encoded explicitly. In this context, user interfaces were reviewed that allow for a synchronized presentation of visual and acoustic music content.
4. The *"Meaning"* group addressed the subject of how musical meaning can be derived from musical data and, in particular, from musical sound. Here, the path from the given low-level (acoustic) raw data to high-level musical models was traced from a human-based as well as from a machine-based perspective.
5. In the *"Ground Truth"* panel, fundamental issues related to the interpretation, usage, and generation of ground truth annotations were discussed. Some of the questions raised during the panel were: What is ground truth in music? How can one handle inter- and intra-annotator variations? How can the quality of ground truth be evaluated? Are there alternatives to manually generated ground truth annotation?
6. Finally, in the *"Grand Challenges"* panel we discussed in which way music information research may and should impact our daily lives and our society in the future. Here fundamental questions were how to provide the best music for each person, how to predict specific effects of music on our society, and how to enrich human relationships by music.

Beside the extensive discussion of these six topics, we had a number of additional contributions where participants presented more specific research results. These contributions covered a number of different topics such as audio parameterization, music alignment and synchronization, singing voice processing, crowd music listening, music tagging, music indexing, interfaces for music exercises, personalization issues in music search, analysis of ethnic music, and many more.

These topics were complemented by some more interdisciplinary contributions relating the field of music processing to neighboring fields such as speech processing, musicology, music perception, and information retrieval. For example, we discussed the ways in which the field of music processing has benefitted from older fields such as speech processing and how music processing might give something back to these fields. Furthermore, a musicologist reported on the difficulties and resistance experienced when introducing novel computer-based methods into traditional humanistic sciences such as musicology. Another highlight of our seminar was a keynote presentation given by Hannah Bast on her CompleteSearch Engine that allows for very fast processing of complex queries on large text collections.

Conclusions

In our seminar, we addressed central and groundbreaking issues on how to process music material given in various forms corresponding to different musical aspects and modalities. In view of the richness and complexity of music, there will be no single strategy that can cope with all facets equally well. Therefore unifying frameworks and fusion approaches are needed which allow for combining, integrating, and fusing the various types of information sources to support music analysis and retrieval applications. Also, to further enhance our field, one needs to understand better the complex relationships within music as well as the complex effects of music on the human mind, thus requiring interdisciplinary research efforts. The Dagstuhl seminar gave us the opportunity for discussing such issues in an inspiring and retreat-like atmosphere. The generation of novel, technically oriented scientific contributions was not the focus of the seminar. Naturally, many of the contributions and discussions were on a rather abstract level, laying the groundwork for future projects and collaborations. Thus the main impact of the seminar is likely to take place in the medium to long term. Some more immediate results, such as plans to share research data and software, also arose from the discussions. As measurable outputs from the seminar, we expect to see several joint papers and applications for funding (e.g. to the European Union) proceeding from the discussions held at Dagstuhl.

Beside the scientific aspect, the social aspect of our seminar was just as important. We had an interdisciplinary, international, and very interactive group of researchers, consisting of leaders and future leaders in our field. Most of our participants visited Dagstuhl for the first time and enthusiastically praised the open and inspiring atmosphere. The group dynamics were excellent with many personal exchanges and common activities. Younger scientists mentioned their appreciation of the opportunity for prolonged discussions with senior researchers—something which is often impossible during conference-like events.

In conclusion, our expectations of the seminar were not only met but exceeded, in particular with respect to networking and community building. Last but not least, we heartily thank the Dagstuhl board for allowing us to organize this seminar, the Dagstuhl office for their great support in the organization process, and the entire Dagstuhl staff for their excellent services during the seminar.

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

Panel Discussions

3 Overview of Talks

3.1 Multimodal Processing in a Digital Music Library Context

David Bainbridge (University of Waikato, NZ)

License Creative Commons BY-NC-ND 3.0 Unported license
© David Bainbridge

In the digital library research group at the University of Waikato we have been (or are in the process of) working with several aspects of multimodal music processing to assist in the formation of digital libraries, and the augmentation of the services they offer. In my presentation, I have addressed the following issues.

- A digital music stand that is fully immersed in the digital library, and whose functionality is enhanced through automated content processing.
- A spatial-hypermedia approach to organizing and managing "musical moments" that occur as part of the process of composition.
- Using digital library collections that combine linked-data with music analysis to capture (and ultimately publish) the outcome of MIR experiments.

3.2 The CompleteSearch Engine

Hannah Bast (Universität Freiburg, DE)

License Creative Commons BY-NC-ND 3.0 Unported license
© Hannah Bast

Joint work of Bast, Hannah; Weber, Ingmar
Main reference Hannah Bast and Ingmar Weber: "The CompleteSearch Engine: Interactive, Efficient, and Towards IR& DB Integration," CIDR 2007, pp. 88–95.

CompleteSearch is a new search engine technology for very fast processing of complex queries on large text collections. Supported query types are: prefix search, faceted search, error-tolerant search, synonym search, database-like search, and semantic search. CompleteSearch is highly interactive: hits are displayed instantly after each keystroke, along with suggestions for possible continuations / refinements of the query. In my talk, I showed exciting demos of the various features of CompleteSearch and also said a few words about their realization.

3.3 What can fMRI tell us about MIR?

Michael Casey (Dartmouth College - Hanover, US)

License Creative Commons BY-NC-ND 3.0 Unported license
© Michael Casey

Music representation requires making decisions about what information to extract, or declare. We outline details of a set of studies that are designed to test the predictive power of different music representations via-a-vis human neural responses in fMRI experiments. Our methods combine audio music feature extraction, latent variable analysis, linguistic and symbolic representations, and multi-variate pattern analysis on neural population codes. To this end, we designed and implemented a Python software framework, named Bregman, built upon OMRAS2 AudioDB C++ framework, that facilitates the design of experiments combining

MIR and fMRI methods. The software is currently being used by a group of 19 graduate and undergraduate students studying music, information, and neuroscience.

3.4 Ground Truth: Where Aggregates Fail

Elaine Chew (USC - Los Angeles, US)

License Creative Commons BY-NC-ND 3.0 Unported license
© Elaine Chew

Differences in composer vs. listener annotations of music structure in improvised performances with the Mimi system was shown, and presented evidence for individual emotion perception variability. The Mimi structure analysis project is joint work with Alexandre François, Isaac Schankler, and Jordan Smith; the latter project is a Master's thesis by Merrick Mosst; both are work done at the MuCoaCo lab at USC.

3.5 Towards Automated Processing of Multimodal Representations of Music

Michael Clausen (Universität Bonn, DE)

License Creative Commons BY-NC-ND 3.0 Unported license
© Michael Clausen

Joint work of Clausen, Michael; Fremerey, Christian; Müller, Meinard
Main reference David Damm, Christian Fremerey, Frank Kurth, Meinard Müller, and Michael Clausen: "Multimodal presentation and browsing of music," Proc. 10th International Conference on Multimodal Interfaces (ICMI), pp. 205–208, 2008.

There are many different types of digital music representations that capture aspects of music on various different levels of abstraction. For example, a piece of music can be described visually by means of sheet music that encodes abstract high-level parameters such as notes, keys, measures, or repeats in a visual form. Because of its explicitness and compactness, most musicologists discuss and analyze the meaning of music on the basis of sheet music. On the other, most people enjoy music by listening to audio recordings, which represent music in an acoustic form. In particular, the nuances and subtleties of musical performances, which are generally not written down in the score, make the music come alive. In this contribution, we discussed various strategies towards automated processing of music data across different representations. In particular, we showed how one can bridge the gap between the sheet music domain and the audio domain discussing aspects on music representations, music synchronization, and optical music recognition, while indicating various strategies and open research problems.

3.6 Multimodality in Human Computer Music Performance

Roger B. Dannenberg (CMU - Pittsburgh, US)

License Creative Commons BY-NC-ND 3.0 Unported license
© Roger B. Dannenberg

Human Computer Music Performance (HCMP) is the integration of computer performers into popular live music. At present, HCMP exists only in very limited forms, due to a lack of understanding of how computer performers might operate in the context of live music and a lack of supporting research and technology. I believe we need to envision a future performance practice that involves computers as musicians. An immediate realization is that non-audio communication is of the utmost importance in real-world music performance. By thinking about HCMP, we discover many interesting multimodal problems. These include: music notation and its use as a two-way communication channel in live performance, musical gestures for conducting and cuing, processing live multi-channel audio, coordination with video and projections in performance, multiples sensors and modalities for beat and measure tracking, and semantic processing concerning intention, emotion, and creation in multiple media.

3.7 Multimodal Music Exercises

Christian Dittmar (Fraunhofer Institute for Digital Media Technology - Ilmenau, DE)

License Creative Commons BY-NC-ND 3.0 Unported license
© Christian Dittmar

URL http://www.songs2see.eu

We presented the project Songs2See that combines multiple music related modalities in order to assist users while learning a musical instrument. The developed software helps budding musicians to do their exercises on the instrument they are acquiring. The concept is similar to music video games; the main difference is the usage of real music instruments instead of game controllers. The most important modalities are the music score and the corresponding music recording, but additional, interactive visualizations supplement the application. We demonstrated the creation of new exercise content by means of semi-automatic music transcription, where backing tracks, which can be used to play along with, are extracted from real-world music recordings. This is achieved using symbolic note transcriptions to initialize the time-varying parameters of a source separation algorithm. At any stage of the content creation process, the user can intervene and correct possibly erroneous transcription results.

3.8 Association and Linking of Related Music Information

Simon Dixon (Queen Mary University of London, GB)

License Creative Commons BY-NC-ND 3.0 Unported license
© Simon Dixon

Many types of music information are available on the Web and in libraries and music databases, including structured and unstructured text (e.g. biographical metadata); complete or partial scores (e.g. chords, tabs, lead sheets, lyrics); and recordings of performances (e.g.

mixed, multi-track, video). For any given work, many such instances might exist, and an ideal music information system would allow navigation between the various representations at both the document and fragment levels, based on the relationships between and within documents that it has discovered. Such a system would enhance both the human experience (e.g. browsing, search) and automatic analysis (e.g. informed transcription, multimodal similarity computation). To realise this vision, the open research questions involve: finding related documents; synchronising them; representing the relationships in ways that are meaningful to users and allow reuse by other systems; and licensing/IP issues.

3.9 Score-Informed Audio Parameterization

Sebastian Ewert (Universität Bonn, DE)

License Creative Commons BY-NC-ND 3.0 Unported license
© Sebastian Ewert
Joint work of Ewert, Sebastian; Müller, Meinard

In recent years, the processing of audio recordings by exploiting musical knowledge as specified by a musical score has turned out to be a promising research direction. Here, one assumes that, additionally to the audio recording to be analyzed, one is given a MIDI file (representing the score) of the same underlying piece of music. The note event information specified by the MIDI file can then be used to support audio analysis tasks such as source separation or instrument equalization. In our contribution, we consider the problem of score-informed audio parameterization with the objective to successively adapt and enrich the note event information provided by the MIDI file to explain the given audio recording. More precisely, our goal is to parameterize the spectrogram of the audio recording by exploiting the score information. Our parameterization approach works iteratively proceeding in several steps. In the first step, we compute a temporal alignment between the MIDI file and the audio recording. Since the alignment accuracy is of major importance, we employ a refined synchronization method that exploits onset information. In the next steps, we successively adapt model parameters referring to dynamics, timbre, and instrumentation such that the spectrogram described by the model reflects the audio spectrogram as accurately as possible.

3.10 Time and Perception in Music and Computation

Alexandre R.J. François (Harvey Mudd College - Claremont, US)

License Creative Commons BY-NC-ND 3.0 Unported license
© Alexandre R.J. François
Main reference Alexandre R.J. François: "Time and Perception in Music and Computation," in New Computational Paradigms for Computer Music, G. Assayag and A. Gerzso (Eds.), Editions Delatour France, IRCAM, pp. 125–146, 2009.

Music has long been a fascinating domain of application for computer scientists. It is a particularly challenging and inspiring one, as it only exists at the confluence of creation, representation and performance. Two "un-computational" properties characterize musical tasks. First, they are geared towards human perception and cognition. Computation aims at super-natural exactness, boring consistency, and absolute reproducibility; on the other hand, perception and cognition operate in the realm of natural variability, exciting unmet expectations, and approximate live reproductions. Secondly, musical tasks epitomize the constant struggle between the desire to stop time and the necessity to live (and experience) in the present; traditional computing paradigms abstract and collapse these two sides of time onto arbitrarily antagonistic sets of compromises. The resulting abstract manifestation of time in computing is enforced as a strong invariant, universally and implicitly relied upon. Hermes/dl, a design language created specifically for the specification of complex dynamic systems, addresses many of the challenges posed by the computational modeling of musical tasks. As a design language, Hermes/dl consists of a collection of primitives, a set of organizing principles, and collections of qualifying situations. A graphical notation shields users from traditional algebraic notation, in an effort to make Hermes/dl more accessible and appealing to potential users in creative and scientific fields.

3.11 Groundtruthing for SALAMI and Billboard Projects

Ichiro Fujinaga (McGill University - Montreal, CA)

License Creative Commons BY-NC-ND 3.0 Unported license
© Ichiro Fujinaga

I presented two recent annotation projects undertaken at McGill University. One of them is a structural analysis of about 1000 pieces of variety of music as part of the Structural Analysis of Large Amount of Musical Information (SALAMI) project. The other one is a harmonic analysis of popular music listed on the Billboard Hot 100 list from 1950 to 1990. We have annotated about 600 songs so far. At the end of my talk, I posed the following questions to the participants: What is ground truth in music? What are the uses of ground truth in music research? Are there different kinds of ground truth? Are there different qualities of ground truth? How does one evaluate the quality of ground truths? Do we know when we have enough amount of ground truth? Should all ground truth be created by human beings? By any human beings? What are the effects of inter- and intra-annotator variability? Can ground truth be generated by machines in certain applications?

3.12 Crowd Music Listening: Internet-Based Music Listening with Shared Semantic Information

Masataka Goto (AIST - Ibaraki, JP)

License Creative Commons BY-NC-ND 3.0 Unported license
© Masataka Goto
URL http://en.wikipedia.org/wiki/Nico_Nico_Douga

I introduced a rapidly-growing video sharing web service in Japan, called "NICO NICO DOUGA." Although this service was not invented by ourselves, I informed the Dagstuhl participants of some advanced features that can enhance music listening experiences, shared interesting phenomena we observed on this service, and discussed whether essential advanced ideas on this service can be spread to other countries/cultures in the future. The NICO NICO DOUGA is a very popular web service managed by a Japanese company, Niwango. This service started in December 2006 and now has 19,610,000 registered users, which is more than 15 % of Japanese citizens. It won the Japanese Good Design Award in 2007, and an Honorary Mention of the Digital Communities category at Prix Ars Electronica 2008. On this service, users can upload, share, and view video clips like YouTube, but it supports networked text communication where comments by anonymous users are overlaid on video clips and are synchronized to a specific playback time. In this way, comments can be related to events in the video. Since it can create a sense of shared watching experience called *"Pseudo-Synchronized Communication"* (coined by Satoshi Hamano), users can feel as if they enjoy together. We observed a lot of interesting ways of using comments and social tags, such as barrage ("DANMAKU"), ASCII art, original/parody lyrics, impressions, interpretations, and feelings. Since those time-synchronous comments can be considered semantic information about video content (music, for example), such internet-based music listening with shared semantic information can enhance our music listening experiences. In my keynote talk at AdMIRe 2009 of IEEE ISM 2009, I coined a new term *"Crowd Music Listening"* for this new way of music listening. You can enjoy music together with the crowd. You are not alone anymore while listening to music.

3.13 A Method for Obtaining Semantic Facets of Music Tags

Fabien Gouyon (INESC Porto, PT)

License Creative Commons BY-NC-ND 3.0 Unported license
© Fabien Gouyon

Main reference Mohamed Sordo, Fabien Gouyon, and Luís Sarmento: "A Method for Obtaining Semantic Facets of Music Tags," Workshop on Music Recommendation and Discovery, ACM Conference on Recommender Systems, Barcelona, 2010.
URL http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-633/wom2010_paper4.pdf

Music folksonomies have an inherent loose and open semantics, which hampers their use in structured browsing and recommendation. In my Dagstuhl talk, I presented a method (detailed in more length in a WOMRAD 2010 paper) for automatically obtaining a set of semantic facets underlying a folksonomy of music tags. The semantic facets are anchored upon the structure of the dynamic repository of universal knowledge Wikipedia. We also illustrated the relevance of the obtained facets for the automatic classification of Last.fm tags.

3.14 Why and How should MIR Research Open up to all Music Modalities?

Fabien Gouyon (INESC Porto, PT)

License Creative Commons BY-NC-ND 3.0 Unported license
© Fabien Gouyon

There are diverse modalities to music, namely, auditory (through hearing), natural language (through hearing, sight), gesture (through sight), physiology (through touch, temperature, and other inner senses), and so on. That is, humans can make sense of, or associate a meaning to a musical phenomenon transmitted via any of the above modalities (e.g. recognize a musical instrument when hearing it, understanding a song's topic by reading its lyrics). Furthermore, there are interactions between modalities (e.g. shivering when hearing a given song). Understanding how humans achieve these associations is certainly an interesting scientific goal, and one of the ways to try and understand this is the computational way: building machines that should do the same. In my Dagstuhl talk, I intended to bring forward the notion that in most "classic" MIR research topics (e.g. genre classification), which often focused on the auditory modality, we may gain valuable insights by considering a multimodal approach. This may be a worthy approach to the understanding of multimodal associations we make when listening to music. But embracing more modalities in our research is certainly not an easy task. There are several fundamental issues including data issues, methodologies issues, and researcher training issues. Regarding data issues, I put forward a few questions:

- How can researchers obtain data from a particular modality? Are there inherent difficulties?
- Which are inherent issues with data cleaning and ground truth annotations for each particular modality?
- How can general data from a given modality help leverage music-specific research in that same modality? For instance, how can the availability of general text (e.g. on Internet) leverage improvements in analyses of music-related textual data (e.g. last.fm tags)?
- How can we relate data from diverse modalities? For instance, is it the best course of research to try mapping low-level audio features (auditory) to written labels of arbitrary high levels of semantics (natural language) via machine learning? Does it even really make sense to try to do so?

In this Dagstuhl talk, I did not intend to answer all the above questions of course, but rather tried to foster discussions on these. I also showed some of the work we do at the Sound and Music Computing group in Porto (http://smc.inescporto.pt), where some of the above questions are—albeit not directly addressed—latent scientific concerns in our research.

3.15 Integrated Content-based Audio Retrieval Framework

Peter Grosche (MPI für Informatik - Saarbrücken, DE)

License 🅭 🄏 ⊜ Creative Commons BY-NC-ND 3.0 Unported license
© Peter Grosche
Joint work of Grosche, Peter; Müller, Meinard

Even though there is a rapidly growing corpus of audio material, there still is a lack of efficient systems for content-based audio retrieval, which allow users to explore and browse through large music collections without relying on manually generated annotations. To account for the various user requirements, we plan to develop a content-based retrieval framework that supplies and integrates various functionalities for flexibly adjusting and tuning the underlying retrieval strategy. Based on the query-by-example paradigm, the user will be able to mark an arbitrary passage within a music representation. This passage is then used as query to retrieve all documents from the music collection containing parts or aspects similar to this passage. The new approach is to develop a framework that facilitates flexible and intuitive control mechanisms for adjusting various aspects in the search process. Firstly, the user may specify the musical properties to be considered in the similarity search. This allows to search for rhythmic, melodic, or harmonic patterns. Secondly, the framework will integrate various retrieval strategies ranging from high-specificity audio identification, over mid-specificity audio matching to low-specificity cover song and genre identification. Here, the goal is to supply the user with a control mechanism that allows to seamlessly adjust the specificity level in the search process. Thirdly, the retrieval framework will provide functionalities that account for subdocument as well as document-level retrieval. In combination with suitable visualization, navigation, and feedback mechanisms, the user is then able to successively refine and adjust the query formulation as well as the retrieval strategy.

3.16 Integrating Different Knowledge Sources for the Computational Modeling of Flamenco Music

Emilia Gómez (Universitat Pompeu Fabra - Barcelona, ES)

License Creative Commons BY-NC-ND 3.0 Unported license
© Emilia Gómez

Main reference Joaquín Mora, Francisco Gómez, Emilia Gómez, Francisco Escobar-Borrego, and José Miguel Díaz-Báñez: "Characterization and Melodic Similarity of A Cappella Flamenco Cantes," Proc. 11th Int'l Society for Music Information Retrieval Conference (ISMIR), pp. 351–356, 2010. Catherine Guastavino, Francisco Gómez, Godfried Toussaint, Fabrice Marandola, and Emilia Gómez: "Measuring Similarity between Flamenco Rhythmic Patterns," Journal of New Music Research 38(2), pp. 129–138, 2009.
URL http://mtg.upf.edu/research/projects/cofla

There is a wealth of literature on music research that focuses on the understanding of music similarity from different viewpoints and on the computation of similarity distances to cluster different pieces according to composer, performer, genre or mood. This similarity measure is often based on comparing musical excerpts in audio format and measuring the distance of a set of content descriptors representative of different musical facets (e.g. the used instruments, rhythmic pattern or harmonic progression). Alternative approaches are based on comparing context information from the contrasted pieces (e.g. influences, temporal and geographical coincidences), which is usually extracted from the web or manually labelled. A combination of these two approaches (content and context) seems to be the most adequate solution, but there is still a limitation on current approaches. This might be due to the fact that there are still other information sources to consider, such as the listening conditions. State-of-the-art research has mainly focused on the analysis of music from the so-called "Western tradition," given that most music retrieval systems are targeted toward this kind of music. Nevertheless, some studies are now considering if the available descriptors and similarity distances are suitable when dealing with music from other traditions. In this situation, the notion of similarity is also affected by the listener's cultural background and his previous exposure to the considered musical structures.

We focus here in the study on flamenco music. Flamenco is a music tradition mostly originally from Andalusia, in southern Spain. The origin and evolution of flamenco styles and variants have been studied by different disciplines, mainly ethnomusicology, anthropology or literature. Prior studies have mainly focused on artists' biographies, lyrics and social context, and there are few works on music analysis. There are some difficulties and motivations for developing computational models of similarity in flamenco music: being an oral tradition, there are no written scores; there exist few quality historical recordings and music collections are spread and not consistently documented; flamenco is not as present on the web as other musical styles; cultural institutions and music platforms are concerned with the preservation and spreading of flamenco music, given its commercial and cultural interest.

The goal of our project is to develop computational models for computer-assisted description, similarity computation, comparative analysis and processing of flamenco music (http://mtg.upf.edu/research/projects/cofla). We want to integrate different knowledge sources: content description (computational analysis of recordings and music similarity algorithms), context information (expert analyses) and user modeling (human judgements). As a case study, we deal with flamenco a capella singing. We focus on melodic similarity, and we evaluate state-of-the-art algorithms for automatic transcription and similarity computation. We work with a music collection of the most representative performances from 4 different a capella singing styles, mainly Debla, Martinete and Toná. We focus on analyzing the melodic exposition, and we approach both inter-style classification and intra-style similarity. Some of the challenges of this project are strongly connected to three of the topics discussed at Dagstuhl Seminar on Multimodal Music Processing: Multimodality, Evaluation and Ground Truth, and User Modeling.

- **Multimodality.** Flamenco research is highly multimodal, as we can combine different inputs when analyzing a certain performance: audio, video, context information (style and singer) and lyrics. These modalities are sometimes complementary and sometimes contradictory. In addition, we do not always have access to all of them simultaneously, so the main research to be done here is related to their integration in current algorithms. As mentioned before, this project intends to combine some of these modalities and their related knowledge sources by means of combined measures of similarity.
- **Evaluation and Ground Truth.** There are some difficulties in this project when evaluating algorithms for two different tasks: melodic transcription and similarity measurement. Regarding melodic transcription, we have adopted a two-stage evaluation procedure (annotations vs. corrections): first, we have collected manual melodic contour transcriptions from flamenco experts, where time information is not relevant and ornaments are removed; then, we have asked them to perform manual corrections and refinements of detailed transcriptions provided by the computational model. This allows us to gather and contrast both annotations of overall melodic contour and ornaments (melisma information). In the same way, we have gathered different sources of ground truth information for music similarity: list of relevant features, similarity ratings and validation of clusters/trees generated by computational models. We have observed that different ground-truth sources are complementary, and there is always a degree of subjectivity in each of them. Computational models should then integrate them and adopt procedures for interactive-validation and user-adapted annotation.
- **User Modeling.** Our project deals with two different user profiles: musicians with little knowledge of flamenco music and experts with high knowledge of flamenco music. We have observed a low correlation among their similarity ratings and the features they use to compare flamenco performances. Here, one particular challenge is to implement user-adapted similarity measures.

In conclusion, the integration of distinct knowledge sources as well as user adaptation are two key aspects in developing a computer-assisted model of flamenco music retrieval.

3.17 Audio Signal Representations for Temporal Structure Segmentation

Florian Kaiser (TU Berlin, DE)

License 🅭 🄏 ⊜ Creative Commons BY-NC-ND 3.0 Unported license
© Florian Kaiser

Music structural segmentation is the core of many MIR applications and remains the focus of many research activities. Most solutions proposed for this task were designed to either detect repetitive patterns or homogeneous acoustical segments in a self-similarity matrix of the audio signal. Depending on the parametrization, it is thus assumed that feature frames extracted over musical sections present some sort of statistical invariance, or that the sequence of features will be exactly repeated while the section is repeated. This assumption is however rarely fulfilled, and similarity matrices are often unable to yield a proper visualization of the actual structure. Modeling features over a local short-time horizon, some work has shown that introducing contextual information in the measure of similarity between two time instants of the audio signal could enhance the structure visualization. We follow this idea and propose to concatenate local chroma sequences in "Multi-Probe Histograms." Transitions between major chroma bins of adjacent frames are mapped to the bins of a histogram. This yields a fixed-size representation of the chroma sequence that condenses local temporal relations between feature frames and summarizes the whole sequence. Preliminary results show that embedding this modeling in a similarity matrix, structure visualization can be strongly enhanced. We would like to discuss the musical interpretation of such representations of chroma sequences, especially with regard to their relation to tonality and harmony.

3.18 Between two Domains: Synergies of Music and Speech Information Retrieval

Frank Kurth (Fraunhofer FKIE, Wachtberg - Bonn, DE)

License 🅭 🄏 ⊜ Creative Commons BY-NC-ND 3.0 Unported license
© Frank Kurth

From its early days on, research in Music Information Retrieval (MIR) has adapted various established technologies from speech processing. Some popular examples are the use of MFCC (Mel Frequency Cepstral Coefficients) features, dynamic time warping or hidden Markov models. Subsequently, those technologies were suitably adopted to the requirements of the music domain, and then considerably extended and complemented by a wide variety of novel methods. In the last years, we were interested in transferring technology in the opposite direction: How can research in Speech Retrieval profit from the—now more mature—methods developed in MIR? In particular, we investigated the domains of feature design, matching techniques, and exploitation of multimodal representations. As a concrete example, we derived novel speech features (HFCC-ENS) from chroma-based music features to develop a method for detecting short sequences of spoken words within a speech recording ("keyphrase spotting"), where the underlying search technology is based on an audio matching approach from MIR. In another example, we adapted a strategy from multimodal MIR to time-align a speech recording with a corresponding textual transcript. In contrast to methods from classical speech processing, the use of MIR-based strategies allowed us to develop methods which work in an unsupervised manner.

3.19 Cross-Disciplinary Perspectives on Music Information Retrieval Challenges

Cynthia Liem (TU Delft, NL)

License 🅭 🄏 ⊜ Creative Commons BY-NC-ND 3.0 Unported license
© Cynthia Liem

Many tasks in Music Information Retrieval research take place at a conceptually high level. As the real-world "non-digital" versions of such tasks involve human understanding and interpretation of music, when computational methods are developed for performing these same tasks, methodological challenges occur. A clear-cut solution to overcoming these challenges has not yet been found, which especially poses problems in the stages of result evaluation and interpretation.

Music Information Retrieval is not the first research field in which such challenges occur. Longer-established fields such as Artificial Intelligence, Computational Linguistics, Content-Based Image Retrieval, Multimedia and Interactive Information Retrieval have faced similar problems. The initial plan was to discuss whether we could identify specific directions from such external neighbouring fields that the Music-IR community could learn and benefit from in its own development. For example, are there any success stories or epic failures from other fields than we can follow or avoid? Are there common pitfalls, or open questions that have not been sufficiently resolved yet in any of those fields? Such considerations would stimulate a cross-disciplinary dialogue and exchange of ideas, and indicate opportunities where Music-IR can take initiative and start to play an exemplary role for neighbouring fields. While in the end, this topic did not become an official main discussion point at the seminar, throughout the whole week it became clear that cross-disciplinary knowledge transfers both to and from the MIR domain will indeed be valuable. Concrete neighbouring fields that were mentioned as being of interest and inspiration included Text Information Retrieval and Speech Recognition, but also Linguistics and Musicology, indicating a broad basis for further discussions.

3.20 Analysis of Guitar Tabs and Chord Sequences

Robert Macrae (Queen Mary University of London, GB)

License 🅭 🅮 ⊜ Creative Commons BY-NC-ND 3.0 Unported license
© Robert Macrae

With over four and a half million tablatures and chord sequences (tabs), the world wide web holds vast quantities of hand annotated scores wrapped up in non-standardised ASCII text files. Theses scores are typically so incomplete, with errors, noise and duplicates that simply filtering out the correct tabs is challenging. Despite this, tabs are by far the most popular means of sharing musical instructions on the Internet. We are interested in developing tools that use text analysis and alignment for the automatic retrieval, interpretation and analysis of such files in order to filter out or recreate the true original music score from the multitude of tabs available. Ultimately it is hoped this work could allow for tab recommendation, tab synthesis, tab synchronisation/score following and integration with other music information for music processing.

3.21 Lyrics-to-Audio Alignment: Methods of Integrating Textual Chord Labels and an Application

Matthias Mauch (AIST - Tsukuba, JP)

License Creative Commons BY-NC-ND 3.0 Unported license
© Matthias Mauch

Joint work of Mauch, Matthias; Fujihara, Hiromasa; Goto, Masataka
Main reference Matthias Mauch, Hiromasa Fujihara, and Masataka Goto: "Lyrics-to-Audio Alignment and Phrase-Level Segmentation Using Incomplete Internet-Style Chord Annotations," Proc. 7th Sound and Music Computing Conference (SMC), pp. 9–16, 2010.
URL http://smcnetwork.org/node/1328

Aligning lyrics to audio is by definition a multimodal task. It has a wide range of applications such as the automatic generation of karaoke scores, song-browsing by lyrics, and the generation of audio thumbnails. Existing methods are restricted to using only lyrics and match them to phoneme features extracted from the audio (usually mel-frequency cepstral coefficients). Our novel idea is to integrate the textual chord information provided in the paired chords-lyrics format known from song books and Internet sites into the inference procedure. By doing so the text–audio bimodality is complemented by a second form of multimodality: timbre–harmony.

We proposed two novel methods that implement our idea. Firstly, assuming that all chords of a song are known, we extended a hidden Markov model (HMM) framework by including chord changes in the Markov chain and an additional audio feature (chroma) in the emission vector. Secondly, for the more realistic case in which some chord information is missing, we presented a method that recovers the missing chord information by exploiting repetition in the song. We conducted experiments with five changing parameters and showed that with accuracies of 87.5% and 76.0%, respectively, both methods perform better, with statistical significance, than the baseline. Furthermore, we demonstrated the "Song Prompter" software system, which acts as a performance assistant by showing horizontally scrolling lyrics and chords in a graphical user interface together with an audio accompaniment consisting of bass and MIDI drums. This application allowed us to show that the automatic alignment is accurate enough to be used in a musical performance.

3.22 Synchronization-based Music Audio/Video Annotation

Meinard Müller (MPI für Informatik - Saarbrücken, DE)

License Creative Commons BY-NC-ND 3.0 Unported license
© Meinard Müller

For general music, there still is a significant gap between the demand of descriptive high-level features and the capability of existing feature extractors to automatically generating them. In particular, the automated extraction of high-level metadata from audio representations such as score parameters, timbre, melodies, instrumentation, or lyrics constitutes an extremely difficult problem with many yet unsolved problems. To bridge this gap, we suggest a knowledge-based approach for metadata extraction by exploiting the fact that one and the same piece of music often exists in several versions on different descriptive levels. For example, one version may encode explicit high-level information (score, lyrics, tablature, MIDI) and another version low-level information (audio, CD recording, video clip). Then the strategy is to use the information given by a high-level version in order to support localization and extraction of corresponding events in the low-level version. As a special case, this approach has been successfully applied to align musical onset times of notes (given by some score representation) with corresponding physical onset times in some audio representation—a process, which can be regarded as an automatic knowledge-based annotation of the audio data. For the future, we will systematically refine and extend such a knowledge-based approach to automatically generate various kinds of metadata annotations and linking structures between music documents including new extraction techniques for higher level semantic descriptors such as harmonic progressions, tablature, lyrics, rhythm patterns, or motivic patterns.

3.23 A Multi-Perspective Analysis of Chord Labeling Procedures

Meinard Müller (MPI für Informatik – Saarbrücken, DE)

License Creative Commons BY-NC-ND 3.0 Unported license
© Meinard Müller
Joint work of Verena Konz; Müller, Meinard
Main reference Verena Konz, Meinard Müller, and Sebastian Ewert: "A Multi-Perspective Evaluation Framework for Chord Recognition," Proc. 11th International Conference on Music Information Retrieval (ISMIR), pp. 9–14, 2010.

The automated extraction of chord labels from audio recordings constitutes a major task in music information retrieval. To evaluate computer-based chord labeling procedures, one requires ground truth annotations for the underlying audio material. However, the manual generation of such annotations on the basis of audio recordings is tedious and time-consuming. On the other hand, trained musicians can easily derive chord labels from symbolic score data. In our contribution, we bridge this gap by describing a procedure that allows for transferring annotations and chord labels from the score domain to the audio domain and vice versa. Using music synchronization techniques, the general idea is to locally warp the annotations of all given data streams onto a common time axis, which then allows for a cross-domain evaluation of the various types of chord labels. As a further contribution, we extend this principle by introducing a multi-perspective evaluation framework for simultaneously comparing chord recognition results over multiple performances of the same piece of music. The revealed inconsistencies in the results do not only indicate limitations of the employed chord labeling strategies but also deepen the understanding of the underlying music material. Our multi-perspective visualization based on a musically meaningful time axis has turned out to be a valuable analysis tool for musicological tasks. In collaboration with musicologists, we are now investigating how recurrent tonal centers of a certain key can be determined automatically within large musical works.

3.24 VocaListener: Synthesis of Human-Like Singing by Using User's Singing and Its Lyrics

Tomoyasu Nakano (AIST – Tsukuba, JP)

License Creative Commons BY-NC-ND 3.0 Unported license
© Tomoyasu Nakano

Joint work of Nakano, Tomoyasu; Goto, Masataka
Main reference Tomoyasu Nakano and Masataka Goto: "VocaListener: A Singing-to-Singing Synthesis System Based on Iterative Parameter Estimation," Proc. 6th Sound and Music Computing Conference (SMC), pp. 343–348, 2009.
URL http://smc2009.smcnetwork.org/programme/pdfs/175.pdf

Since 2007, many Japanese users have started to use commercial singing synthesis systems to produce music, and the number of listeners who enjoy synthesized singing is increasing. Over 100,000 copies of singing synthesis software based on Yamaha's Vocaloid have been sold, and various compact discs that include synthesized vocal tracks have appeared on commercial music charts in Japan. We presented the singing synthesis system "VocaListener" that iteratively estimates parameters for singing synthesis software from a user's singing voice with the help of song lyrics. Since a natural voice is provided by the user, the synthesized singing voice mimicking it can be human-like and natural without time-consuming manual adjustments. The iterative estimation provides robustness with respect to different singing synthesis systems and singer databases. Moreover, VocaListener has a highly accurate lyrics-to-singing synchronization function, and its interface lets a user easily correct synchronization errors by simply pointing them out. In addition, VocaListener has a function to improve synthesized singing as if the user's singing skills were improved. Demonstration videos including examples of synthesized singing are available at http://staff.aist.go.jp/t.nakano/VocaListener/. In this seminar, we discussed mechanisms used in the VocaListener, presented some synthesized results, and indicated some future directions. Here, one main goal is to be able to synthesize singing voices that can not be distinguished any longer from human singing voices.

3.25 Modeling Novelty in Music Retrieval

Nicola Orio (University of Padova, IT)

License Creative Commons BY-NC-ND 3.0 Unported license
© Nicola Orio

One interesting aspect of available search engines is that sometimes, or maybe quite often, they fail and retrieve documents that are only slightly related to the user information needs. So they unwillingly introduce novelty in the retrieved list of documents by presenting unexpected results. These results may promote serendipity, help the user to discover new things, and may enlarge user interest—possibly even modifying the initial information needs. This contribution describes a draft idea, far to be tested, on how game theory results may be applied to model and somehow to control novelty in a music retrieval task. In particular, Prisoner's Dilemma seems to provide an interesting approach that is worth being explored. The idea is that parameters, algorithms, and different modalities behave as players who, at each time step, may choose whether or not to collaborate. The goal of each player is to maximize his or her payoff, while the goal of the envisaged system is to encourage players to change their strategy over time, to promote unexpected results to be presented to the user, and to hopefully improve the user's appreciation. Multimodal music retrieval is naturally based on the concept of cooperation between subsystems. Yet, cooperation is a complex phenomenon that encompasses also the possibility that individual strategies are based on betraying, at least from time to time, from which the overall system may be able to retrieve original items.

3.26 Towards a Definition of the Description of Annotated M.I.R. Corpora

Geoffroy Peeters (Ircam - Paris, FR)

License Creative Commons BY-NC-ND 3.0 Unported license
© Geoffroy Peeters

This presentation concerned a proposal for the definition of the description of annotated M.I.R. corpora. Given that the availability of annotated data within a given M.I.R. task usually corresponds to the start of a growing number of research activities (this has been the case for music genre classification, chord recognition, or music structure analysis), accessibility to annotated data must be considered as a major issue. Today, annotation data is often provided by various research labs (or companies) each one using its own annotation methodology and own concept definitions. This is not a problem by itself. However, the lack of descriptions of these methodologies or concepts (What is actually described? How is the data annotated?) is a problem in view of sustainability, usability, and sharing of such corpora. Therefore, it is essential to exactly define what and how annotations of M.I.R. corpa should be supplied and described. It should be noted that usual "standards" such as MPEG-7, RDF or Music-XML only concern the storage of the annotation, but not the description of the annotations themselves. The following issues become crucial for the description of corpora.

- What concepts are actually annotated? For example, in the case of structure annotations, which kind of structure is considered? In the case of chord annotations, what instruments or voices are considered? Only accompaniment, the leading voice, or both? For most M.I.R. corpora, the definitions of the concepts that are actually annotated are missing.
- How is the data actually annotated? Is the data derived from meta-data catalogues, from parameters of a synthesis process, or from human annotations? In the latter case, how many people were involved in the annotation process? What is the reliability of the data? Has cross-validation been applied?
- What are the actual annotation rules? For example, what is the temporal precision used for segment annotations? Which type of dictionary has been used for the labels? Are there equivalences between labels?
- How does the corpus look like? For example, what are the properties of the audio being described? Does the corpus consist of real or synthesized audio recordings? For which purposes have specific music-tracks been selected? Are these artificial recordings made specifically for the purpose of an annotation task?
- What are the storage formats for the raw data and annotation data? For example, in the case of audio material, are the files encoded as PCM or in some compressed form? Was the audio data synthesized from MIDI files? Are the annotations given in MPEG-7, RDF, or Music-XML?

3.27 “Copy and Scale” Method for Doing Time-Localized M.I.R. Estimation: Application to Beat-tracking

Geoffroy Peeters (Ircam - Paris, FR)

License Creative Commons BY-NC-ND 3.0 Unported license
© Geoffroy Peeters

Main reference Geoffroy Peeters: “ ‘Copy and Scale’ Method for Doing Time-Localized M.I.R. Estimation: Application to Beat-tracking.” Journal of New Music Research, Special issue on Music and Machine Learning, to appear.

This presentation concerned a new algorithm, named “copy and scale” algorithm, with the goal to first estimate the local parameters of an unknown item (such as beat positions) simply by locating the closest item in a database and then to copy and scale the annotations of the located item to serve as the estimations for the unknown item. A nearest neighbour algorithm consists in assigning the information (such as music genre or mood) of the closest item of a pre-annotated database to an unknown target. It can be viewed as a “copy and paste” method. It is usually used for estimating global variables (such as genre or mood). The “copy and scale” method we propose allows for estimating local variables (i.e. variables that have a specific time location) and consists in “scaling” the closest item to adapt it to the properties of the unknown target. For this, we first represent the content of an audio signal using a sampled and tempo-normalized complex DFT of an onset-energy-function. This representation is used as the code over which the nearest neighbour (NN) search is performed. Along each code of the NN space, we store the corresponding annotated beat-marker positions in a normalized form. Once the closest code is found in the database, its tempo is assigned to the unknown item and the normalized beat-markers are scaled to this tempo in order to provide the estimation of the unknown item beat-markers. We perform a preliminary evaluation of this method and show that, with such a simple method, we can achieve results comparable to the ones obtained with sophisticated approaches.

3.28 Toward Reverted Indexing for Multimodal Music Retrieval

Jeremy Pickens (Catalyst Repository Systems, US)

License Creative Commons BY-NC-ND 3.0 Unported license
© Jeremy Pickens

Joint work of Pickens, Jeremy; Cooper, Matthew; Golovchinsky, Gene
Main reference Jeremy Pickens, Matthew Cooper, and Gene Golovchinsky: “Reverted Indexing for Feedback and Expansion,” Proc. 19th ACM International Conference on Information and Knowledge Management (CIKM), pp. 26–30, 2010.
URL http://www.fxpal.com/publications/FXPAL-PR-10-581.pdf

Traditional text interactive information retrieval systems operate by creating inverted lists, or term indexes. For every term in the vocabulary, a list is created that contains the documents in which that term occurs and its relative frequency within each document. Retrieval algorithms then use these term frequencies alongside other collection statistics to identify the matching documents for a query. Recent advances turn the process around: instead of indexing documents, we index query result sets. First, a set of basis queries, representative of the collection as a whole, are chosen. Each query is then run through a retrieval system and the resulting document IDs are treated as terms while the score or rank of the document is used as the frequency statistic. Thus, an index of documents retrieved by basis queries is created. We call this index a reverted index. With reverted indexes, standard retrieval algorithms can retrieve the best basis queries (as results) for a set of documents (used as queries). These recovered queries can then be used to identify additional related documents or to aid the user in query formulation, selection, and feedback.

In this presentation, we proposed applying the idea of reverted indexing to multimodal music retrieval by extending the set of basis queries to multiple modalities: pitch, rhythm, tempo, harmony, timbre, lyric, and structure to name but a few. As it is the results of each of these query modalities that are being indexed rather than the features and structures that go into ranking the results of each modality, basis queries from different modalities can be combined into a single index. At runtime, standard retrieval algorithms can again be used to retrieve the best basis queries (as results) for a set of documents (used as queries). These multimodal basis queries can then be recombined in an ad hoc manner to identify related pieces of music.

3.29 Future of Music Search and Consumptions—Results of the CHORUS+ Think Tank at MIDEM2011

Andreas Rauber (TU Wien, AT)

License Creative Commons BY-NC-ND 3.0 Unported license
© Andreas Rauber

CHORUS+ (EU FP7 Coordinating Action) organized a highly focused Think Tank on "Music Search and Consumption" with a small group of key industry experts to identify needs and directions for research in Music IR at MIDEM (Marché international de l'édition musicale). The presentation reported on the key suggestions and findings from this Think Tank, specifically focusing on the gaps between the vision for music search and consumption, what research is working on, and what the market is (planning) to deliver. The Think Tank was supported by an on-line questionnaire, which has also been answered by numerous participants in this seminar. The first part of it is rather standard statistical information while the questions towards the end try to identify the biggest gaps and visions for required tools, technologies and services. Results of both the questionnaire as well as the Think Tank highlight the importance, particularly of personalization and recommender technology, with some disagreement concerning the suitability of existing methods. Further highlights include an anticipated shift from ownership-based access to music to access-based, with business models likely shifting from an object-based towards an attention-based model.

3.30 Multimodal/Multi-level Integration Strategies for Music Signal Processing

Gaël Richard (TELECOM-Paristech – Paris, FR)

License Creative Commons BY-NC-ND 3.0 Unported license
© Gaël Richard

Main reference Olivier Gillet, Slim Essid, and Gaël Richard: "On the Correlation of Audio and Visual Segmentations of Music Videos," IEEE Trans. on Circuits and Systems for Video Technology, 17 (2), pp 347-355, 2007.

The purpose of this talk was to highlight the interest and challenges of multimodality or multi-level integration by discussing a selection of our prior studies in multimodal music signal processing. The first challenge addressed consisted in a discussion on which ways audio source separation could be improved when some additional information about the sources such as the musical score, beats, or the sung melody is known. A potential application of active listening involving audio-informed source separation and watermarking was also discussed.

As second topic, we discussed possible strategies for combining features that are calculated at different frame rates or that represent different semantics levels. In particular, it was shown that, besides the well known early and late fusion strategies, it may be appropriate to consider other approaches that iteratively (or jointly) exploit information given by the various modalities to improve estimates for the other modalities. In the same spirit, an example of cross-modality search was discussed. More specifically, it was shown how cross-correlation between semantic functions automatically extracted from each stream can lead to efficient inter-modality search for information retrieval or content authoring. Here, specific examples on music videos were given including a short demonstration on an audio-based video search prototype.

As a third topic, the strategy of considering the time evolution of audio features simultaneously at different temporal levels was addressed. This so-called multi-level feature integration was discussed with a specific focus on the appropriateness to combine both early and late fusion strategies. Finally, a set of questions was selected intended as initial input for further group discussions.

3.31 Personalization in Multimodal Music Retrieval

Markus Schedl (Universität Linz, AT)

License 🅭 🄏 ⊜ Creative Commons BY-NC-ND 3.0 Unported license
© Markus Schedl

This talk provided an overview of current research endeavors and existing solutions in multimodal music retrieval, where the term “multimodal” relates to two aspects. Accounting for the context of a piece of music or an artist constitutes the first aspect, while the second one relates to incorporating the user context. Adding “traditional” signal-based audio analysis, I proposed a model that incorporates three broad categories of influence factors for music retrieval and music similarity computation: music content, music context, and user context. The music context is introduced as all the information that is important to the music, albeit not directly extractable from the audio signal. Such information includes, for example, editorial or collaboratively assembled meta-data, lyrics in textual form, cultural background of an artist, or images of album covers. The user context, in contrast, is defined by various external factors that influence how a listener perceives music. It is therefore strongly related to user modeling and personalization, both facets of music information research that have not gained large attention by the MIR community so far. In my estimation, adding personalization aspects to existing music retrieval systems (for example, playlist generators, recommender systems, or visual browsers) constitutes one key issue in future MIR research.

3.32 Next Gen Music Analysis: Some Inspirations from Speech

Björn Schuller (TU München, DE)

License Creative Commons BY-NC-ND 3.0 Unported license
© Björn Schuller

Numerous ideas have been transferred in the past between the fields of Music Information Retrieval (MIR) and Automatic Speech Processing (ASP) including the usage of synthesized speech/chords for model training, feature brute-forcing, language modelling, bag-of-words, and Mel-Frequency Cepstral Coefficients. In this light, opportunities for further cross-discipline methodology transfer were discussed starting from a unified perspective on speech and music analysis. Such may include next generation MIR tasks such as recognition of age, height, weight, ethnicity, voice quality, likability, personality, as well as verification of performers and sung language identification. Concerning databases, rich transcription including several MIR tasks for single databases and added meta-information allow for combined analysis. Non-prototypical instance selection and partitioning including development sets are further to become more standard in MIR. Next, many classification and regression approaches recently evolved for ASP including tailored Graphical Models, Long-Short-Term-Memory enhanced architectures, and Gaussian Mixture Model supervectors for Support Vector Machines. Multitask Learning, Grid Search or similar optimization techniques and Self-Learning are further candidates for interesting advances in the field of MIR. Considering Multimodal Music Processing (MMP), fusion by synthesis and asynchronous stream processing provide elaborate techniques used in Audio-visual Speech Recognition. In addition, more fully automated MIR studies could be seen, where all data has to be retrieved from the web or derived from mid-level features based on automatic recognition. Ground truth interpretation by kappa and alpha measures and result interpretation by significances or effect power is also yet to become more common in MIR. Overall, several options for transferring recent developments in speech analysis to the MIR field exists. In the future, transfer from the fields of image and video analysis looking in particular at MMP as well as combined music and speech (and sound) analysis to maximally exploit mutual dependencies may become increasingly interesting.

3.33 PROBADO: A Multimodal Music Library Systems

Verena Thomas (Universität Bonn, DE)

License Creative Commons BY-NC-ND 3.0 Unported license
© Verena Thomas
Joint work of Thomas, Verena; Damm, David; Fremerey, Christian; Clausen, Michael
URL http://www.probado.de

As part of the PROBADO (Prototypischer Betrieb allgemeiner Dokumente) project, we aim at developing a prototypical digital music library system which is tested and implemented at the Bavarian State Library in Munich. Besides an appealing user interface following the WYSIWYH ("What You See Is What You Hear") paradigm, a widely automated processing workflow including digitization, indexing, annotation, and presentation has been developed. To allow for content based search (audio, lyrics, and symbolic representations) as well as convenient document presentation anb browsing, several state-of-the-art MIR techniques have been integrated into the system.

3.34 A User Interface for Motivic Analysis

Verena Thomas (Universität Bonn, DE)

License Ⓒ Ⓢ Ⓔ Creative Commons BY-NC-ND 3.0 Unported license
© Verena Thomas
Joint work of Thomas, Verena; Clausen, Michael

Besides developing methods for identifying patterns in symbolic music documents and determining repetitions of these patterns within the document, it is important to provide a user interface for visualizing the analysis results. We presented a first prototype of such a system offering a piano roll like visualization of the analyzed piece of music. The system identifies repeating patterns within the piece of music, their occurrences throughout the piece as well as the type of occurrence such repetitions with equal musical intervals, retrogrades, inversions and retrograde inversions of musical intervals, as well as repetitions with respect to the note durations. In addition, the hierarchical structure of the identified patterns is analyzed and visualized by the system.

3.35 Mobile Multimodal Music Processing for Edutainment and eHealth

Ye Wang (National University of Singapore, SG)

License Ⓒ Ⓢ Ⓔ Creative Commons BY-NC-ND 3.0 Unported license
© Ye Wang

Multimodal smartphones are powerful enough for real-time music processing and can be easily connected to the Internet wirelessly. In this context, I presented our research endeavors at National University of Singapore, leveraging multimodal music processing on smartphones for edutainment and eHealth applications. We have also built a private cloud computing cluster dedicated for the backend computing and Internet scale MIR.

Edutainment is an important research topic, as it seeks to find new ways for applying technology to the education environment in an entertaining format. Our efforts in this area have produced the MOGCLASS prototype which has been successfully deployed in three primary schools and the Muscular Dystrophy Association (Singapore) with positive feedback. MOGCLASS has received enthusiastic comments from members in the Curriculum Planning and Development Department of the Ministry of Education (Singapore) and has started commercial licensing to schools. e-Health seeks to apply web-based and mobile technologies to create new health solutions. We have begun core research in building systems for elderly-care applications that mix mobile computing and domain specific music retrieval. We want to mention two specific current projects.

- In the MusicalWalk project, we are developing a web based system that allows music therapists to find suitable pieces of music to facilitate rhythmic auditory stimulation (RAS)-based gait training for a patient with Parkinson's disease.
- In the EZSleep project, we are investigating technologies for selecting the best sequence of music for improved sleep quality of an insomniac by learning his or her biophysical indicators and mapping them to content-based musical descriptors.

3.36 What is it Like to be a Musicologist?

Frans Wiering (Utrecht University, NL)

License Creative Commons BY-NC-ND 3.0 Unported license
© Frans Wiering

In this presentation, I investigated the relationship between present-day musicology and computational music processing. These two fields seem to be far apart: computing has at best a role as tool supplier in mainstream musicological research and musicology figures mainly as a source of domain knowledge and ground truth in computational music processing. Setting aside the matter of different methodologies, there seem to exist completely different views as to what music is. One important reason why this is the case relates to the crisis that took place in musicology in the late 1980s.

Before that time, musicology used to be first of all about Western classical music. Its central concept was the musical "work," the autonomous artistic creation exemplified by the score that captured the composer's intention. This "positivist" view of music was not so much superseded as stripped of its credibility in the following decades of "critical" musicology. The work concept itself was shown to be a cultural construction, loaded with prejudice against other possible views of music. Critical musicology has today become mainstream musicology, examining musical meaning, subjectivity, culture, context, gender and identity rather than beauty, structure and indeed "the music itself."

Music processing is first of all about music as data, whether notation, audio or otherwise. Its benefits are the well-known ones of computation: size, speed, accuracy and formal modelling, but the mismatch with musicology is obvious. However, a common ground may be found in what David Huron has called a "new empiricist" approach. Studies of context, subjectivity and meaning generation may benefit from data-rich approaches complementing the anecdotic evidence that they generally are based on. For example, in meaning generation high-level musical patterns shared between different pieces play an important role, but to understand this role massive amounts of music need to be analysed.

3.37 What does it Mean to "Process" "Music"?

Geraint A. Wiggins (Goldsmiths, University of London, GB)

License Creative Commons BY-NC-ND 3.0 Unported license
© Geraint A. Wiggins

There is a significant gap (sometimes called a "semantic" gap, but that is only partly correct in one particularly narrow usage of the word "semantic") between the audio signal, which corresponds with "Sound" in the air, and "Music" which is the experience that arises in an appropriately enculturated human mind as a result of the application of an appropriate audio stimulus. My research is about understanding "Music" from this viewpoint, and identifying representations that properly model the human experience of music, and studying their properties. This entails a certain amount of perceptual and cognitive psychology, since, I claim, the word "Music" has no meaning except where a mind is involved. In this my presentation, I tried to bridge the (non-semantic) gap between audio signals and the human experience of music, via what are sometimes called symbolic, mid-level, or discrete representations. While none of these names are incorrect, none of them captures the point: to be meaningful, the representations need to be perceptually and/or cognitively valid.

3.38 Music Commentator: A System for Generating Comments Synchronized with Music Audio Signals

Kazuyoshi Yoshii (AIST – Ibaraki, JP)

License Creative Commons BY-NC-ND 3.0 Unported license
© Kazuyoshi Yoshii

Joint work of Yoshii, Kazuyoshi; Goto, Masataka
Main reference Kazuyoshi Yoshii and Masataka Goto: "MusicCommentator: Generating Comments Synchronized with Musical Audio Signals by a Joint Probabilistic Model of Acoustic and Textual Features," Lecture Notes in Computer Science, Vol. 5709, pp. 85–97, 2009.
URL http://dx.doi.org/10.1007/978-3-642-04052-8_8

In online video sharing services, many users often provide free-form text comments for temporal events in music clips, not for entire clips. Although this is one of the most interesting functions of humans, its mechanism has not fully been investigated so far. How can computers emulate the commenting behavior of humans? We have developed a system called "Music Commentator" that suggests possible natural-language comments on appropriate temporal positions in a musical audio clip. To achieve this, we propose a joint probabilistic model of audio signals and text comments. The model is trained by using existing clips and users' comments given to those clips. Given a new clip and some of its comments, the model is used to estimate what temporal positions could be commented on and what comments could be added to those positions. It then concatenates possible words by taking language constraints into account. Our experimental results showed that using existing comments in a new clip resulted in improved accuracy for generating suitable comments to it. To improve the system, we have to take into account high-level semantic musical features related to melody, rhythm, and harmony. In addition, visual features of music video clips should be dealt with.

3.39 A Statistical Approach to the Ground-Truth Problem

Kazuyoshi Yoshii (AIST – Ibaraki, JP)

License Creative Commons BY-NC-ND 3.0 Unported license
© Kazuyoshi Yoshii

So far, humans have put a lot of effort to annotate musical pieces. Such annotations are used as ground-truth data and have been greatly contributing to the progress of music information research. Note that some kinds of annotations are forced to follow humans' arbitrary decisions. For example, how finely should we categorize genres, chords, or moods? There is no way to measure the appropriateness of such decisions in a principled manner. To solve this problem, we could use a statistical "data-driven" approach. Bayesian nonparametrics is considered to be especially promising for opening the door to a meta-level research area, i.e., a new generation of computational musicology.

4 Working Groups

4.1 Group Session: Signal Models for and Fusion of Multimodal Information

Sebastian Ewert, Masataka Goto, Peter Grosche, Florian Kaiser, Kazuyoshi Yoshii, Frank Kurth, Matthias Mauch, Meinard Müller, Geoffroy Peeters, Gaël Richard, Björn Schuller

License Creative Commons BY-NC-ND 3.0 Unported license
© Sebastian Ewert, Masataka Goto, Peter Grosche, Florian Kaiser, Kazuyoshi Yoshii, Frank Kurth, Matthias Mauch, Meinard Müller, Geoffroy Peeters, Gaël Richard, Björn Schuller

The *"Model"* group discussed the issue of how signal models can be developed that exploit multimodal information. In most music processing methods a piece of music is analyzed by extracting high-level semantics from a single source of information, for example from an audio recording or a symbolic score representation (MIDI, MusicXML). However, there might be additional data sources available such as corresponding videos, chord transcriptions, lyrics annotations, tags and text comments, or other textual meta-data. Given the complexity of most analysis tasks, it is a very promising idea to combine all available sources of information to support music analysis methods or to validate their results. During the group meeting we discussed several aspects related to this idea.

In particular, we exemplarily selected several music analysis approaches that incorporate multiple information sources, and discussed them in detail with the goal to identify general fusion strategies. On the one hand, we categorized strategies into either early or late fusion approaches and discussed their advantages and weaknesses. Here, the terms *early* and *late* refer to the stage in the processing pipeline where the sources of information are finally combined. On the other hand, we discussed aspects related to the modeling of uncertainty involved in the combination process. Here, uncertainty may arise from missing or contradictory information in the given data.

Overall, we identified several general fusion strategies. Firstly, we discussed the *weighting-of-streams* approach where streams of information are processed individually and the results are merged using a weighting scheme. Secondly, we identified the *iterative-reinforcement approach* where unreliable or coarse information is refined in an iterative manner. Furthermore, we discussed how uncertainty is treated in classical machine learning methods. Here, we considered *confusion aware approaches*, where information about the reliability of individual classifiers is gathered from training data in a first step and is used later to build better classifiers. As one major issue, we found that a measure of uncertainty involved in the combination process is a prerequisite for choosing a suitable fusion strategy.

4.2 Group Session: User-aware Music Information Retrieval

Nicola Orio, Emilia Gómez, Fabien Gouyon, Cynthia Liem, Tomoyasu Nakano, Jeremy Pickens, Andreas Rauber, Markus Schedl, Ye Wang

License Creative Commons BY-NC-ND 3.0 Unported license
© Nicola Orio, Emilia Gómez, Fabien Gouyon, Cynthia Liem, Tomoyasu Nakano, Jeremy Pickens, Andreas Rauber, Markus Schedl, Ye Wang

The *"User"* group addressed the topic of user-aware music information retrieval. Actually, the perception of music is highly subjective and depends on the personal background and taste. Furthermore, the perception also depends on the specific user context, which is determined by various external factors that influence how a listener perceives music in a certain situation. In this session, we discussed the central question on how contextual information can be integrated into the retrieval process in order to account for long-term user behavior and short-term user interests. In particular, the consideration of personalization aspects in the development of music retrieval systems has been identified as one key issue in future MIR research. Additionally, we also discussed how search engines may yield satisfying results in terms of novelty, popularity, and serendipity of the retrieved items. Here, we conjectured that retrieval results that are only slightly related to the user information needs may actually become interesting for promoting serendipity helping the user to discover new and surprising music.

4.3 Group Session: Symbolic Music Representations and OMR

Christopher Raphael, Ichiro Fujinaga, Simon Dixon, Robert Macrae, David Bainbridge, Michael Clausen, Verena Thomas

License Creative Commons BY-NC-ND 3.0 Unported license
© Christopher Raphael, Ichiro Fujinaga, Simon Dixon, Robert Macrae, David Bainbridge, Michael Clausen, Verena Thomas

In the session of the *"Symbol"* group, we discussed the question of how to bridge the gap between visual, symbolic, and acoustic representations of music. Our discussion of symbolic music representations largely centered around optical music recognition (OMR) as this is the most obvious possibility for getting large quantities of symbolic music data for classical music. We feel the need for OMR especially now, since the International Music Score Library Project offers a very large open library of classical music scores. Several of the group members, Fujinaga and Bainbrige have considerable experience working in the problem, though current technology is not adequate for many commonly encountered situations. We discussed particular ways to pose the problem that allow for useful results with only a partial mastery of the grand challenge.

Much of our discussion focused on the SyncPlayer work of Clausen and Thomas (and others) which registers orchestral audio with scores and allows for simultaneously viewing and listening. All were very much hoping to see this project succeed and become widely available (beyond the University of Bonn and the Bavarian State Library). There still remains quite a bit of work before this can be achieved, including solving both technical issues and intellectual property concerns. We also discussed other symbolic representations such as guitar tabs, which are widely used by guitarists and others interested in a chord/lyric description of a song. There remain many interested challenges here, including synchronization with audio and video material.

4.4 Group Session: Meaning of Music

Geraint A. Wiggins, Frans Wiering, Michael Casey, Elaine Chew, Roger B. Dannenberg, Alexandre R.J. François, Matthias Mauch

License Creative Commons BY-NC-ND 3.0 Unported license
© Geraint A. Wiggins, Frans Wiering, Michael Casey, Elaine Chew, Roger B. Dannenberg, Alexandre R.J. François, Matthias Mauch

The *"Meaning"* group addressed the subject of how musical meaning can be derived from musical data and, in particular, from musical sound. Here, we addressed the gap between the audio signal, which corresponds with "Sound" in the air, and "Music" which is the experience that arises in an appropriately enculturated human mind as a result of the application of an appropriate audio stimulus. To better understand this gap, we sketched a path from the given low-level (acoustic) raw data to high-level musical models from a human-based as well as from a machine-based perspective.

5 Panel Discussions

5.1 Panel Session: Ground Truth

Ichiro Fujinaga, Geoffroy Peeters, Kazuyoshi Yoshii, Meinard Müller, Elaine Chew

License Creative Commons BY-NC-ND 3.0 Unported license
© Ichiro Fujinaga, Geoffroy Peeters, Kazuyoshi Yoshii, Meinard Müller, Elaine Chew

In the panel session on *"Ground Truth"* fundamental issues related to the interpretation, usage, and generation of ground truth annotations were discussed. What is ground truth in music? How can one handle inter- and intra-annotator variabilities? How can the quality of ground truth be evaluated? Are there alternatives to manually generate ground truth annotations? These are just some of the questions raised during the panel. There were five presentation interspersed with lively discussions. The presentations were made by Ichiro Fujinaga ("Groundtruthing for SALAMI and Billboard Projects," see Section 3.11), Geoffroy Peeters ("Towards a Definition of the Description of Annotated M.I.R. Corpora," see Section 3.26), Kazuyoshi Yoshii ("A statistical approach to the ground-truth problem," see Section 3.39), Meinard Müller ("A Multi-Perspective Analysis of Chord Labeling Procedures," see Section 3.23), and Elaine Chew ("Ground Truth: Where Aggregates Fail," see Section 3.4).

5.2 Panel Session: Grand Challenges

Masataka Goto, Roger B. Dannenberg

License Creative Commons BY-NC-ND 3.0 Unported license
© Masataka Goto, Roger B. Dannenberg

Main reference Masataka Goto: "Music Informatics," IPSJ Magazine, Vol. 51, No. 6, pp. 661–668, 2010.
URL http://staff.aist.go.jp/m.goto/PAPER/IPSJ201006goto.pdf

In the panel session *"Grand Challenges"* we discussed in which way music information research may and should impact our daily lives and our society in the future. Among others, the following challenges were discussed:

- How to provide the best music for each person?

- How to predict music trends?
- How to enrich the relation between humans and music?
- How to push new music evolution forward?
- How to contribute to solving environmental and energy problems?
- How to involve computers in live performances?
- How to predict specific effects of music?

There were two presentations by Masataka Goto and Roger Dannenberg ("Multimodality in Human Computer Music Performance," see Section 3.6), which were accompanied by some passionate discussions between and within the audience and the panelists.

Participants

- David Bainbridge
  University of Waikato, NZ
- Hannah Bast
  Universität Freiburg, DE
- Michael Casey
  Dartmouth College - Hanover, US
- Elaine Chew
  USC - Los Angeles, US
- Michael Clausen
  Universität Bonn, DE
- Roger B. Dannenberg
  CMU - Pittsburgh, US
- Christian Dittmar
  Fraunhofer Institut IDMT - Illmenau, DE
- Simon Dixon
  Queen Mary University of London, GB
- Sebastian Ewert
  Universität Bonn, DE
- Alexandre R. J. Francois
  Harvey Mudd College - Claremont, US
- Ichiro Fujinaga
  McGill University - Montreal, CA
- Emilia Gómez
  Universitat Pompeu Fabra - Barcelona, ES
- Masataka Goto
  AIST - Ibaraki, JP
- Fabien Gouyon
  INESC Porto, PT
- Peter Grosche
  MPI für Informatik - Saarbrücken, DE
- Florian Kaiser
  TU Berlin, DE
- Frank Kurth
  Fraunhofer FKIE, Wachtberg - Bonn, DE
- Cynthia Liem
  TU Delft, NL
- Robert Macrae
  Queen Mary University of London, GB
- Matthias Mauch
  AIST - Tsukuba, JP
- Meinard Müller
  Saarland University and MPI für Informatik - Saarbrücken, DE
- Tomoyasu Nakano
  AIST - Tsukuba, JP
- Nicola Orio
  University of Padova, IT
- Geoffroy Peeters
  Ircam - Paris, FR
- Jeremy Pickens
  Catalyst Repository Systems, US
- Christopher Raphael
  Indiana University - Bloomington, US
- Andreas Rauber
  TU Wien, AT
- Gael Richard
  TELECOM-Paristech - Paris, FR
- Markus Schedl
  Universität Linz, AT
- Björn Schuller
  TU München, DE
- Verena Thomas
  Universität Bonn, DE
- Ye Wang
  National University of Singapore, SG
- Frans Wiering
  Utrecht University, NL
- Geraint A. Wiggins
  Goldsmiths, University of London, GB
- Kazuyoshi Yoshii
  AIST - Ibaraki, JP