D.2.2: Written Report on MIR state-of-the-art

Grant Agreement nr	287711
Project acronym	MIReS
Start date of project (dur.)	Oct 1st 2011 (18 months)
Document due Date :	June 2012
Actual date of delivery	30 June 2012
Leader	INESC Porto
Reply to	fgouyon@inescporto.pt
Document status	review integrated, v.2

Project funded by ICT-7th Framework Program from the European Commission

Project ref. no.	287711
Project acronym	MIReS
Project full title	Roadmap for Music Information ReSearch
Document name	D2.2 Written report on MIR state-of-the-art
Security (distribution level)	PP
Contractual date of delivery	June 2012
Actual date of delivery	30 June 2012
Deliverable name	D2.2 Written report on MIR state-of-the-art
Type	R
Status & version	review integrated, v.2
Number of pages	
WP / Task responsible	WP2, INESC Porto
Other contributors	All partners
Author(s)	Fabien Gouyon
EC Project Officer	Rossella Magli
Abstract	
Keywords	
Sent to peer reviewer	Yes (IRCAM)
Peer review completed	Yes (IRCAM)
Circulated to partners	not yet
Read by partners	not yet
Mgt. Board approval	not yet

A BACKGROUND

This deliverable is a companion document to deliverable D3.1 on "Specification of the Roadmapping Process" and the MIReS Roadmap itself. It reports on reviewing activities by members of the consortium in order to provide an in-depth, critical overview of past and current research trends in MIR.
This document reports on the initial review of MIR research trends, carried out during the first 9 months of the project. This review may still suffer changes until the end of the project and the final updated version will be an integral part of the Roadmap itself (i.e. D3.3).

B INTRODUCTION

This deliverable reports on collaborative efforts made by the consortium to review the state-of-the-art in MIR research. As reported in deliverable D3.2, the structure of the Roadmap and the roadmapping process were defined in two consortium meetings at month 1 and month 4. During these meetings, the main topics of MIR research were identified. The structure of the roadmap logically results from this: each topic is described by a concise state-of-the-art and current and future challenges for this topic. In the remainder of this document, we summarize these topics (Section 3), provide the current status of the state-of-the-art (Section 4) and conclude on next steps to follow (Section 5).

C STRUCTURING THE STATE-OF-THE-ART REVIEW

Since the inception of the project, the primary objective of the MIReS review was to provide an in-depth, critical overview of past and current research trends in MIR. More particularly, the consortium aimed at a critical reflection on the evolution of the field and on achievements and pitfalls from its beginning to today, both in terms of research findings but also employed methodologies, societal impacts and exchanges -inward and outward- with neighboring fields of science.

With this aim in mind, the consortium therefore decided to focus the review of MIR state-of-the-art on three different perspectives: (1) a technical-scientific perspective, (2) a socio-cultural perspective and (3) an exploitation --or application-based-- perspective. These were then broken down in the following topics:

- Technical-scientific perspective
  - Musically relevant data
  - Music representations
  - Data processing methodologies
  - Knowledge-driven methodologies
  - Music content analysis
  - Interface and interaction aspects
  - Evaluation of research results
- Socio-cultural perspective
  - Social aspects
  - Culture specificity
  - User behaviour
- Exploitation perspective
  - Music Industry Applications
  - Artistic applications
  - Research and educational applications
  - Creative industries applications

The initial content of the review will be published online on the project Wiki (http://mires.eecs.qmul.ac.uk/wiki/index.php/Main_Page) and will serve as "bootstrap" for further revisions by the community at large.

D STATE-OF-THE-ART

The following text compiles all of the State-of-the-Art writing in one place. The text forms integral part of D3.2: Intermediary version of the roadmap. Reviewers may wish to refer to the text contained here, or read it as part of D3.2, where it is linked to the relevant challenges.

1. Introduction: the MIReS vision

The field of Music Information Retrieval (MIR) has centered primarily on the analysis of sound signal for the purpose of more efficient search and faster access to digital collections of recorded music. By expanding its context and addressing challenges such as multimodal information, multiculturalism and multidisciplinarity, MIR has the potential for a major impact on the future economy, the arts and education, not merely through applications of technical components, but also by evolving to address questions of fundamental human understanding, with a view to building a digital economy founded on "uncopiable intangibles": personalisation, interpretation, embodiment, findability and community. Within this wider context we propose to refer to the field of MIR as Music Information ReSearch (MIReS) and thus widen its scope, ensuring its focus is centered on quality of experience with greater relevance to human networks and communities.

1.1 Definition of MIReS

Music Information Research (MIReS) covers all the research topics involved in the understanding and modeling of music by using information processing methodologies. These research topics can either be viewed by technical-scientific, sociocultural or exploitation perspectives. Each perspective promotes different and complementary research approaches and solutions to the overall problem.

2. Technical-scientific perspective

Music Information Research (MIReS) comprises the research aimed at processing music related digital data with the goal to develop musical applications. This includes the issues about the organization of the available digital data about music, the development of new methodologies to process and understand that data, and the development of technologies for specific music applications that take advantage of the data processing methodologies.

2.1 Musically relevant data

We define “musically relevant data” as any type of machine readable data that can be analysed by algorithms and that can give us relevant information for the development of musical applications. We are concerned with both the data that is already available and the research necessary for gathering new relevant data.

2.1.1 State of the art

Music Information Research (MIReS) is so far to a large degree concerned with music only, neglecting the many other forms of media where music also plays an important role. Maybe still ten years ago the main media concerned with music were of course audio itself on CDs and played on terrestrial radio, music videos on TV and printed text in music magazines. Today music seems to be an all-encompassing experience that is an important part of videos, computer games, Web applications, mobile apps and services, artistic applications, etc. In addition to printed text on music there exists a vast range of websites, blogs and specialized communities caring and publishing about music. It will be an important next step for Music Information Research to come to terms with this new broad multi-modality and leave its music-only "niche". Therefore it is necessary for MIReS to broaden its horizon and include a multitude of yet untapped data sources in its research agenda.

Consequently, we try to answer the question what data exists that could be of interest for the general goals of MIReS but that has not yet been exploited by our research community. In doing so, we will present a systematic overview of data already in use plus a listing of sources of data that have been largely overlooked so far.

Data that is available for Music Information Research can be categorized into three different subgroups [Schedl & Knees 2011]: (i) audio-content which is any kind of information computed directly from the audio signal; (ii) music-context is all information relevant to music which is not directly computable from the audio itself like e.g. cover artwork, lyrics, but also artist's background and collaborative tags connected to the music; (iii) user-context is any kind of data that allows us to model a single user in one specific usage setting. User-context data reaches from user's interactions with and feedback to a recommendation system to a user's social and spatio-temporal context all the way to a user's momentary mood and physiological parameters. In addition, we will dedicate a section to the non-trivial task of collecting all these different kinds of music-related data.

Audio content

Let us start with the most prevalent source of data: audio content, i.e. any kind of information computed directly from the audio. Such information is commonly referred to as "features", with a certain consensus on distinguishing between low-level and high-level features (see e.g. [Casey et al. 2008]). Please see section 2.2 (music representations) and 2.5 (music content analysis) for an overview concerning different kinds of features. It is obvious that audio content data is by far the most widely used and researched form of information in our community. This can e.g. be seen by looking at the tasks in last year's "Music Information Retrieval Evaluation eXchange" (MIREX 2011, http://www.music-ir.org/mirex/wiki/2011:Main_Page). MIREX is the foremost yearly community-based framework for formal evaluation of MIR algorithms and systems. Out of the 16 tasks, all but one (Symbolic Melodic Similarity) deal with audio analysis including challenges like: Audio Classification, Cover Song Identification, Audio Key Detection to Structural Segmentation and Audio Tempo Estimation.

Concerning the availability of audio content data there are several legal and copyright issues. Just to give an example, the by far largest data set in MIR, the “Million Songs Dataset” (http://labrosa.ee.columbia.edu/millionsong) does not include any audio, only the derived features. In case researchers need to compute their own features they have to use services like “7-Digital” to access the audio. Collections that do contain the audio as well are usually very small like e.g. the well known “GTzan” collection assembled by George Tzanetakis in 2002 consisting of 1000 songs freely available from the Marsyas webpage (http://marsyas.info/download/data_sets). The largest freely downloadable audio data set is the “1517 Artists” collection introduced by Klaus Seyerlehner (http://www.seyerlehner.info) consisting of 3180 songs from 1517 artists. There also exist alternative collaborative databases of Creative Commons Licensed sounds like Freesound (http://

www.freesound.org/). More on the legal issues concerning availability of audio data is provided in section 4.1.1.3.

Music context

Music context is all information relevant to a music item under consideration that cannot be extracted from the respective audio file itself (see e.g. [Schedl & Knees 2009] for an overview). Just to give some examples, the music video corresponding to a song, the geographical origin of an artist or their socio-cultural background are all additional information that has a decisive impact on how a piece of music is perceived by its recipient.

Symbolic: An important source of information is of course symbolic data. That includes the score of a piece of music if it is notated, but also MIDI, Music XML, sequencer data or in general all kinds of abstract representations of music. Such abstract representations of music can be very close to audio content like e.g. the score to one specific audio rendering but they are usually not fully isomorphic. Therefore we file symbolic data under music context. Going beyond more traditional annotations, recent work in MIR [Macrae & Dixon 2011] turned its attention to tablatures and chord sequences, which are a form of hand annotated scores in non-standardised text files. These tabs are probably the most popular means of sharing musical instructions on the internet (e.g. www.ultimate-guitar.com alone contains more than 2.5 million guitar tabs). At the first MIR conference (ismir2000.ismir.net) a large part of the contributed papers were concerned with symbolical data. A paper on the "Past, Present and Future" of MIR [Uitdenbogerd et al 2000] noted that "interesting developments have occurred in the field of audio retrieval", which probably tells a lot about the status of audio-based MIR in the year 2000. Almost ten years later this imbalance seems to have reversed with authors [Downie et al 2009] lamenting that "ISMIR must rebalance the portfolio of music information types with which it engages" and that "research exploiting the symbolic aspects of music information has not thrived under ISMIR". Symbolic annotations of music present legal and copyright issues just like real audio (see section 4.1.1.3), but collections for e.g. MIDI do exist (http://www.free-midi.org).

Web: A large part of research on music context is strongly related to web content mining. Over the last decade, mining the World Wide Web has been established as another major source of music related information. Music related data mined from the Web can be distinguished into "editorial" and "cultural" data. Whereas editorial data originates from music experts and editors often associated with music labels, cultural data makes use of the wisdom of the crowd by mining large numbers of music related websites including social networks. Advantages of web based music information retrieval are the vast amount of available data as well as its potential to access high-level semantic descriptions and subjective aspects of music not obtainable from mere audio based analysis alone. E.g. it is possible to construct term profiles from artist-related Web pages to derive music similarity information. RSS feeds can be extracted and analyzed or playlists (e.g., radio stations and mix tapes, i.e., user-generated playlists) and Peer-to-Peer networks are other valuable sources of information. Very often co-occurrence analysis is commonly employed to derive similarity on the artist- or track-level. Co-occurrences of artist names on Web pages are also used to infer artist similarity information and for artist-to-genre classification. Song lyrics as a source of music context-related information are analyzed to derive similarity information, e.g. for mood and genre classification. Another source for the music context are collaborative tags, mined for example from last.fm [Levy & Sandler 2007] or gathered via tagging games [Turnbull et al 2007]. There are a number of yet unsolved challenges to tap into the full potential of Web based music analysis. Information obtained automatically from the Web is inherently noisy and erroneous which requires special techniques and care for data clean-up. Data about new and lesser known artists in the so-called "long tail" is usually very sparse which introduces an unwanted popularity bias [Celma 2010]. A list of data sets frequently used in Web-based MIR is provided by Markus Schedl (http://www.cp.jku.at/people/schedl/datasets.html). The "Million Songs Dataset" (h ttp://labrosa.ee.columbia.edu/millionsong) contains some web-related information like e.g. tag information provided by Last.fm.

Video: A possibly very rich source of additional information on music content that has so far received little attention in our community is music videos. The most prominent source for music videos is of course YouTube (www.youtube.com), but alternatives like Vimeo (www.vimeo.com) exist. YouTube is a video-sharing website where registered users can upload and share videos and anyone is allowed to watch these videos free of charge. Although the uploaded material contains anything from amateur clips to video blogs to complete movies, a large part of it are music videos. Whereas a lot of the content on YouTube has been uploaded by individuals which may entail all kinds of copyright and legal issues, some large media companies have lately decided to also offer some of their content. There exists a lively community around the so-called TRECVid campaign (trecvid.nist.gov), a forum, framework and conference series on video retrieval evaluation, much like MIREX (www.music-ir.org/mirex) in our community. One of the major tasks in video information retrieval is automatic labelling of videos, e.g. according to genre, which can be done either globally or locally [Brezeale & Cook 2008]. Typical information extracted from videos are visual descriptors like color, its entropy and variance, hue, or a range of temporal cues like cuts, fades, dissolves. Object-based feautures like the occurence of faces or text and motion-based information like motion density and camera movement are also of interest. Text-based information derived from sub-titles, transcripts of dialogues, synopsis or user tags is another valuable source. A potentially very promising approach seems to be the combined analysis of a music video and its corresponding audio, pooling information from both the image and audio signals. This should add extra information for a whole range of audio tagging tasks (e.g. genre, mood, instrument recognition) as well as for video labelling tasks. Combination of general audio and video information is an established topic in the literature, see e.g. [Wang et al 2003] for an early survey. There already is a limited amount of research explicitly on music videos exploiting both the visual and audio domain [Gillet et al 2007].. A rare and rather specialized example from the MIR community is a recent work on automatically identifying guitar chords using audio and video of the performer [Hrybyk & Kim 2010]. Although the TRECVid evaluation framework (http://trecvid.nist.gov) supports a "Multimedia event detection evaluation track" consisting of both audio and video, to our knowledge no data set dedicated specifically to music videos exists.

Another yet untapped source is books on musicology that are part of Google Books (http://books.google.com/). Google books is a search engine that searches the full text of books if they have already been scanned and digitized by Google. This offers the possibility of using computers to analyse text books on music thereby introducing MIR topics to the new emerging field of digital humanities.

User context

As stated above, user-context data is any kind of data that allows us to model a single user in one specific usage setting. In most MIR research and applications so far, the prospective user is seen as a generic being for whom a generic one-for-all solution is sufficient. Typical systems aim at modeling a supposedly objective music similarity function which then drives music recommendation, play-listing and other related services. This however neglects the very subjective nature of music experience and perception. Not only do different people perceive music in different ways depending on their likes, dislikes and listening history, but even one and the same person will exhibit changing tastes and preferences depending on a wide range of factors: time of day, social situation, current mood, location, etc. Personalizing music services can therefore be seen as an important topic of future MIR research.

Following recent proposals [Goeker & Myrhaug 2002, Schedl & Kness 2011], we like to distinguish five different kinds of user context data: (i) Environment Context, (ii) Personal Context, (iii) Task Context, (iv) Social Context, (v) Spatio-temporal Context. The environmental context is defined as all entities that can be measured from the surrounding of a user, like other people and things, climate including temperature and humidity, noise and light. The personal context can be divided into the physiological context and the mental context. Whereas physiological context refers to attributes like weight, blood pressure, pulse, or eye color, the mental context is any data describing a user's psychological aspects like stress level, mood, or expertise. The task content should describe all current activities pursued by the user including actions and activities like e.g. direct user input to smart mobile phones and applications, but also interaction with diverse messenger and microblogging services. The latter is a valuable source for a user's social context giving information about relatives, friends, or collaborators. The spatio-temporal context reveals information about a user's location, place, direction, speed, and time. As a general remark, the recent emergence of "always on" devices (like e.g. smart phones) equipped not only with a permanent Web connection, but also with various built-in sensors, has remarkably facilitated the logging of user context data from a technical perspective. Integrated GPS modules, accelerometers, light and noise sensors as well as interfaces to almost every Web 2.0 service makes user context logging easier than ever before providing data for all context categories described above. Data sets on the user context are still very rare but e.g. the "user - song - play count triplets" and the Last.fm tags of the "Million Songs Dataset" (http://labrosa.ee.columbia.edu/millionsong) could be said to contain personalized information.

Collecting music related data

For the research /training involved in the development of MIR algorithms as well as for benchmarking them obtaining relevant data for the description of the music content is a major issue. In [Peeters and Fort, 2012] an overview of the different practices of annotated MIR corpora is proposed. Currently, several methodologies are used for collecting these data: - creating an artificial corpus [Yeh et al., 2007], recording corpuses [Goto, 2006] or sampling the world of music according to specific criteria (Isophonics [Mauch et al., 2009], Salami [Smith et al., 2011], Billboard [Burgoyne et al., 2011], MillionSong [Bertin-Mahieux et al., 2011]). The data can then be obtained using experts (this is the usual manual annotation [Mauch et al., 2009]), using crowd-sourcing [Levy, 2011] or so-called games with a purpose (Listen-Game [Turnbull et al., 2007], TagATune [Law et al., 2007], MajorMiner [Mandel and Ellis, 2008]) or by aggregating other contents (Guitar-Tab [McVicar and De Bie, 2010] MusicXMatch, LastFM in the case of the MillionSong). As opposed to other domains, micro-working (such as Amazon Mechanical Turk) is not (yet) a common practice in the MIR field. These various methodologies for collecting data involve various costs: from the most expensive (traditional manual annotation) to the less expensive (aggregation or crowd-sourcing). They also involve various qualities of data. This is related to the inter-annotator and intra-annotator agreement which is rarely assessed in the case of MIR. Compared to other fields, such as NLP or speech, music-related data collection or creation does not follow dedicated protocols. One of the major issues in the MIR field will be to better define protocols to make reliable annotated MIR corpus. Another important aspect is how our research community relates itself to initiatives aiming at unifying data formats in the world wide web. Initiatives that come to mind are e.g. linked data (http://linkeddata.org) which is a collection of of best practices for publishing and connecting structured data on the Web and, especially relevant for MIR, MusicBrainz (http://musicbrainz.org/) which strives to become the ultimate source of music information or even the universal lingua franca of music.

References

- [Bertin-Mahieux et al., 2011] Bertin-Mahieux, T., Ellis, D., Whitman, B., and Lamere, P. (2011). The million song dataset. In Proc. of ISMIR, Miami, USA.
- [Brezeale & Cook 2008] D. Brezeale and D.J. Cook. Automatic Video Classification: A Survey of the Literature. IEEE Transactions on Systems, Man, and Cybernetics, Part C, Volume 38, Number 3, pp. 416-430, 2008.
- [Burgoyne et al., 2011] Burgoyne, J. A., Wild, J., and Ichiro, F. (2011). An expert ground-truth set for audio chord recognition and music analysis. In Proc. of ISMIR, Miami, USA.
- [Casey et al. 2008] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-Based Music Information Retrieval: Current Directions and Future Challenges. Proceedings of the IEEE, 96:668-696, April 2008.
- [Celma 2010] O. Celma. Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space. Springer-Verlag New York Inc, 2010.
- [Downie et al 2009] S.J. Downie, D. Byrd, T. Crawford. Ten Years of ISMIR: Reflections On Challenges and Opportunities. In Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009), Kobe, Japan, 2009.
- [Gillet et al 2007] O. Gillet, S. Essid and G. Richard, On the Correlation of Audio and Visual Segmentations of Music Videos. IEEE Transactions on Circuits and Systems for Video Technology, 17 (2), March 2007, pp 347-355.
- [Goeker & Myrhaug 2002] A. Goeker and H. I. Myrhaug. User Context and Personalisation. In Proceedings of the 6th European Conference on Case Based Reasoning (ECCBR 2002): Workshop on Case Based Reasoning and Personalization, Aberdeen, Scotland, September 2002.
- [Goto, 2006] Goto, M. (2006). Aist annotation for the rwc music database. In Proc. of ISMIR, pages pp.359–360, Victoria, Canada.
- [Hrybyk & Kim 2010] A. Hrybyk, Y. Kim. Combined Audio and Video Analysis for Guitar Chord Identification. In Proceedings 12th International Conference on Music Information Retrieval (ISMIR 2010), pp. 159-164, 2010.
- [Law et al., 2007] Law, E. L. M., Ahn, L. v., Dannenberg, R., and Crawford, M. (2007). Tagatune: A game for music and sound annotation. In Proc. of ISMIR, Vienna, Austria.
- [Levy & Sandler 2007] M. Levy and M. Sandler. A semantic space for music derived from social tags. In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007), Vienna, Austria, September 2007.
- [Levy, 2011] Levy, M. (2011). Improving perceptual tempo estimation with crowd-sourced annotations. In Proc. of ISMIR, Miami, USA.

- [Macrae & Dixon 2011] R. Macrae and S. Dixon. Guitar Tab Mining, Analysis and Ranking. Proceedings of the 12th International Society for Music Information Retrieval Conference, pp 453-458, 2011.
- [Mandel and Ellis, 2008] Mandel, M. I. and Ellis, D. P. W. (2008). A web-based game for collecting music metadata. In Journal of New Music Research, volume 37, pages 151–165. Taylor & Francis.
- [Mauch et al., 2009] Mauch, M., Cannam, C., Davies, M., Dixon, S., Harte, C., Kolozali, S., Tidhar, D., and Sandler, M. (2009). OMRAS2 metadata project 2009. In Proc. of ISMIR (Late-Breaking News), Kobe, Japan.
- [McVicar and De Bie, 2010] McVicar, M. and De Bie, T. (2010). Enhancing chord recognition accuracy using web resources. In Proceedings of 3rd inter- national workshop on Machine learning and music, pages 41–44. ACM.
- [Peeters and Fort, 2012] Peeters, G. and Fort, K. (2012). Towards a (better) definition of the description of annotated m.i.r. corpora. In Proc. of ISMIR, Porto, Portugal.
- [Schedl & Knees 2011] M. Schedl and P. Knees. Personalization in Multimodal Music Retrieval. Proceedings of the 9th International Workshop on Adaptive Multimedia Retrieval (AMR 2011), Barcelona, Spain, July 18-19, 2011.
- [Schedl & Knees 2009] M. Schedl and P. Knees. Context-based Music Similarity Estimation. In Proceedings of the 3rd International Workshop on Learning the Semantics of Audio Signals (LSAS 2009), Graz, Austria, December 2009.
- [Smith et al., 2011] Smith, J. B. L., Burgoyne, J. A., Fujinaga, I., De Roure, D., and Downie, J. S. (2011). Design and creation of a large-scale database of structural annotations. In Proc. of ISMIR, Miami, USA.
- [Turnbull et al 2007] D. Turnbull, R. Liu, L. Barrington, and G. Lanckriet. A Game-based Approach for Collecting Semantic Annotations of Music. In Proceedings of the 8th Interna- tional Conference on Music Information Retrieval (ISMIR 2007), Vienna, Austria, September 2007.
- [Uitdenbogerd et al 2000] A.L. Uitdenbogerd, A. Chattaraj, J. Zobel: Music IR: Past, Present and Future. Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR 2000), Plymouth, Massachusetts October 23-25, 2000.
- [Wang et al 2003] H. Wang, A. Divakaran, A. Vetro, S.-F. Chang, H. Sun. Survey of Compressed-Domain Features used in Audio-Visual Indexing and Analysis. Journal of Visual Communication and Image Representation 14(2), 150-183, 2003.
- [Yeh et al., 2007] Yeh, C., Bogaards, N., and Roebel, A. (2007). Synthesized polyphonic music database with verifiable ground truth for multiple f0 estimation. In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR07), pages 393–398.

2.2 Music representations

Alongside the availability of data, another important issue is the way in which data is (or is not) structured, both at the conceptual level (data representation) and the implementation level (data file formats). Issues of particular relevance for music information research include: the choice of audio features, which are often categorised by level of abstraction into high, mid, and low-level features; use of symbolic languages, ontologies, taxonomies and folksonomies for structuring music information; and means of visualisation and sonification of music-related data.

2.2.1 State of the art

While audio recordings represent musical performances with a high level of detail, there is no direct relationship between the individual samples and the experience of music, which involves notes, beats, instruments, phrases or melodies (the musicological perspective), and which might give rise to memories or emotions associated with times, places or events where identical or similar music was heard (the user perspective). Although there is a large body of research investigating the relationship between music and its meaning from the philosophical and psychological perspectives [e.g., Minsky, 1981; Robinson, 1997; Cross and Tolbert, 2008; JMM], scientific research has tended to focus more on bridging the "semantic gap" between audio recordings and the abstractions that are found in various types of musical scores, such as pitches, rhythms, melodies and harmonies. This work is known as semantic audio or audio content analysis (see section 2.5).

In order to simplify the extraction of useful information from audio recordings, a standard practice is to compute intermediate representations at various levels of abstraction. At each level, features can describe an instant in time (e.g. the onset time of a note), a segment or time interval (e.g. the duration of a chord) or the whole piece (e.g. the key of a piece). Various sets of features and methods for evaluating their appropriateness have been catalogued in the MIR literature [McKinney and Breebaart, 2003; Peeters, 2004; Kim et al., 2005; McEnnis et al., 2005; Pachet and Roy, 2007].

Low-level features relate directly to signal properties and are computed according to simple formulae. Examples are the zero-crossing rate, spectral centroid and global energy of the signal. Time-domain features such as the amplitude envelope and attack time are computed without any frequency transform being applied to the signal, whereas spectral features such as centroid, spread, flatness, skewness, kurtosis and slope require a frequency transform such as the short time Fourier transform (STFT) or constant-Q transform (CQT) [Brown 1991] to be applied as a first processing step.

Mid-level features (e.g. pitches and onset times of notes) are characterised by more complex computations, where the algorithms employed are not always successful at producing the intended results. Typically a modelling step will be performed (e.g. sinusoidal modelling), and the choice of parameters for the model will influence results. For example, in Spectral Modelling Synthesis [Serra and Smith, 1990], the signal is explained in terms of sinusoidal partials tracks created by tracking spectral peaks across analysis frames, plus a residual signal which contains the non-sinusoidal content. The thresholds and rules used to select and group the spectral peaks determine the amount of the signal which is interpreted as sinusoidal. This flexibility means that the representation with respect to such a model is not unique, and the optimal choice of parameters is dependent on the task for which the representation will be used.

High-level features (e.g. genre, tonality, rhythm, harmony and mood) correspond to the terms and concepts used by musicians or listeners to describe aspects of music. To generate such features, the models employed tend to be more complex, and might include a classifier trained on a relevant data set, or a probabilistic model such as a hidden Markov model (HMM) or dynamic Bayesian network (DBN). Automatic extraction of high-level features is not reliable, which means that in practice there is a tradeoff between the expressiveness of the features (e.g. number of classes they describe) and the accuracy of the feature computation.

It should also be noted that the the classification of features into categories such as "high-level" is not an absolute judgement, and some shift in usage is apparent, resulting from the search for ever higher levels of abstraction in signal descriptors. Thus features which might have been described as high-level a decade ago might now be considered to be mid-level features. Also features are sometimes described in terms of the models used to compute them, such as psychoacoustic features (e.g. roughness, loudness and sharpness) which are based on auditory models. Some features have been standardised, e.g. in the MPEG7 standard [Kim et al., 2005]. Another form of standardisation is the use of ontologies to capture the semantics of data representations and to support automatic reasoning about features, such as the Audio Feature Ontology proposed by Fazekas [2010].

In addition to the literature discussing feature design for various MIR tasks, another strand of research investigates the automatic generation of features [e.g., Pachet and Roy, 2007, 2009]. This is a pragmatic approach to feature generation, whereby features are generated from combinations of simple operators and tested on the training data in order to select suitable features. The approach has been shown to be superior to the use of standard feature sets for classification tasks.

Much music information is not in the form of audio recordings, but rather symbolic representations of the pitch, timing, dynamics and/or instrumentation of each of the notes. There are various ways such a representation can arise. First, via the composition process, for example when music notation software is employed, a score can be created for instructing the musicians how to perform the piece. Alternatively, a score might be created via a process of transcription (automatic or manual) of a musical performance. For electronic music, the programming or performance using a sequencer or synthesiser could result in an explicit or implicit score. For example, electronic dance music can be generated, recorded, edited and mixed in the digital domain using audio editing, synthesis and sequencing software, and in this case the software's own internal data format(s) can be considered to be an implicit score representation.

In each of these cases the description (or prescription) of the notes played might be complete or incomplete. In the Western classical tradition, it is understood that performers have a certain degree of freedom in creating their rendition of a composition, which may involve the choice of tempo, dynamics and articulation, or also ornamentation and sometimes even the notes to be played for an entire section of a piece (an improvised cadenza). Likewise in Western pop and jazz music, a work is often described in terms of a sequence of chord symbols, the melody and the lyrics; the parts of each instrument are then rehearsed or improvised according to the intended style of the music. In these cases, the resulting score can be considered to be an abstract representation of the underlying musical work. However not all styles of music are based on the traditional Western score. For example, freely improvised and non-Western musics might have no score before a performance and no established language for describing the performance after the fact.

A further type of music information is textual data, which includes both structured data such as catalogue metadata and unstructured data such as music reviews and tags associated with recordings by listeners. Structured metadata might describe the composers, performers, musical works, dates and places of recordings, instrumentation, as well as key, tempo, and onset times of individual notes. Digital libraries use metadata standards such as Dublin Core and models such as the Functional Requirements for

Bibliographic Records (FRBR) to organise catalogue and bibliographic databases. To assist interoperability between data formats and promote the possibility of automatic inference from music metadata, ontologies have been developed such as the Music Ontology [Raimond et al., 2007].

Looking beyond the conceptual organisation of the data, we briefly address its organisation into specific file formats, and the development and maintenance of software to read, write and translate between these formats. For audio data, two types of representations are used: uncompressed and compressed. Uncompressed (or pulse code modulated, PCM) data consists of just the audio samples for each channel, usually prepended by a short header which specifies basic metadata such as the file format, sampling rate, word size and number of channels. Compression algorithms convert the audio samples into model parameters which describe each block of audio, and these parameters are stored instead of the audio samples, again with a header containing basic metadata. Common audio file formats such as WAV, which is usually associated with PCM data, provide a package allowing a large variety of audio representations. Standard open source software libraries such as libsndfile are available for reading and writing common non-proprietary formats, but some file formats are difficult to support with open source software due to the license required to implement an encoder.

For symbolic music data, a popular file format is MIDI (musical instrument digital interface), but this is limited in expressiveness and scope, as it was originally designed for keyboard instrument sequencing. For scores, a richer format such as MusicXML is required, which includes information such as note spelling and layout is required. For guitar “tabs” (a generic term covering tablature as well as chord symbols with or without lyrics), free text is still commonly used, with no standard format, although software has been developed which can parse the majority of such files [Macrae and Dixon 2011]. Some tab web sites have developed their own formats using HTML or XML for markup of the text files. Other text formats such as the MuseData and Humdrum kern format [Selfridge-Field, 1997] have been used extensively for musicological analysis of corpuses of scores.

For structured metadata, formats such as XML are commonly used, and in particular semantic web formats for linked data such as RDFa, RDF/XML, N3 and Turtle are employed. Since these are intended as machine-readable formats rather than for human consumption, the particular format chosen is less important than the underlying ontology which provides the semantics for the data.

Finally, although music exists primarily in the auditory domain, there is a long tradition of representing music in various graphical formats. Common Western music notation is a primary example, but piano-roll notation and spectrograms also present musical information in a potentially useful format. Since music is a time-based phenomenon, it is common to plot the evolution of musical parameters as a function of time, such as tempo and dynamics curves, which have been used extensively in performance research [Desain and Honing, 1991]. Simultaneous representations of two or more parameters have been achieved using animation, for example the Performance Worm [Dixon et al., 2002], which shows the temporal evolution of tempo and loudness as a trajectory in a two-dimensional space.

References

- [Brown, 1991] J.C. Brown (1991). Calculation of a Constant-Q Spectral Transform, Journal of the Acoustical Society of America, 89 (1), 425-434.

- [Cross and Tolbert, 2008] I. Cross and E. Tolbert (2008). Music and Meaning, in: The Oxford Handbook of Music Psychology (S. Hallam, I. Cross and M. Thaut, Eds.), Oxford University Press.
- [Desain and Honing, 1991] P. Desain and H. Honing (1991). Tempo Curves Considered Harmful: A Critical Review of the Representation of Timing in Computer Music. Proceedings of the International Computer Music Conference, 143-149.
- [Dixon et al., 2002] S. Dixon, W. Goebl and G. Widmer (2002). The Performance Worm: Real Time Visualisation of Expression Based on Langner's Tempo-Loudness Animation. Proceedings of the International Computer Music Conference, 361-364.
- [Fazekas, 2010] G. Fazekas, Audio Features Ontology, http://www.omras2.org/AudioFeatures
- [JMM] The Journal of Music and Meaning. URL: www.musicandmeaning.net
- [Kim et al., 2005] H.G. Kim, N. Moreau and T. Sikora (2005). MPEG7 Audio and Beyond: Audio Content Indexing and Retrieval. Wiley and Sons.
- [Macrae & Dixon 2011] R. Macrae and S. Dixon (2011). Guitar Tab Mining, Analysis and Ranking. Proceedings of the 12th International Society for Music Information Retrieval Conference, pp 453-458.
- [McEnnis et al., 2005] D. McEnnis, C. McKay, I. Fujinaga and P. Depalle (2005). JAudio: A Feature Extraction Library, Sixth International Conference on Music Information Retrieval.
- [McKinney and Breebaart, 2003] M. F. McKinney and J. Breebaart. Features for Audio and Music Classification, Fourth International Conference on Music Information Retrieval.
- [Minsky, 1981] M. Minsky (1981). Music, Mind, and Meaning. Computer Music Journal, Vol. 5, Number 3.
- [Pachet and Roy, 2007] F. Pachet and P. Roy (2007). Exploring Billions of audio features. Proceedings of the 5th International Workshop on Content-Based Multimedia Indexing (CBMI'07).
- [Pachet and Roy, 2009] F. Pachet and P. Roy (2009). Analytical Features: A Knowledge-Based Approach to Audio Feature Generation. EURASIP Journal on Audio, Speech, and Music Processing, 2009:1–23.
- [Peeters, 2004]. G. Peeters (2004). A Large Set of Audio Features for Sound Description, http://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf
- [Raimond et al., 2007] Y. Raimond, S. Abdallah, M. Sandler and F. Giasson (2007). The Music Ontology. Proceedings of the 8th International Conference on Music Information Retrieval, 417-422.
- [Robinson, 1997] J. Robinson (ed., 1997). Music and Meaning. Cornell University Press.
- [Selfridge-Field, 1997] E. Selfridge-Field (ed., 1997) Beyond MIDI: The Handbook of Musical Codes. MIT Press.
- [Serra and Smith, 1990] X. Serra and J. Smith (1990). "Spectral Modeling Synthesis: A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition," Computer Music Journal, 14 (4), 12-24.

2.3 Data processing methodologies

Since its origins, the MIR community has been focusing mainly on bottom-up approaches. For one thing, this is reflected in the fact that MIR is a data-intensive field of research, as illustrated in Sections 2.1 and 2.2. Further, methodologies used in MIR, and consequent algorithms developed, have also been strongly influenced by bottom-up approaches coming from other fields of science. The purpose of this Section is to briefly report on that influence. On the other hand, the relatively smaller influence of top-down methodologies, as found in other fields of science is the purpose of Section 2.4 below.

2.3.1 State of the art

The origins of Music Information Research were multi-disciplinary in nature. At the first edition of the ISMIR conference series, in 2000 (http://ismir2000.ismir.net/), although the number of research papers was significantly smaller than in later editions, papers drew ideas from a relatively large number of disciplines: Digital libraries, Information Retrieval, Musicology and symbolic music analysis, Music representation, Speech processing, Signal processing, Perception and cognition, Image processing (with applications to Optical Music Recognition), User modeling. This initial conference also debated Intellectual Property matters and systematic evaluations.

Since then, the ISMIR conference has grown tremendously, as illustrated by the number of unique authors that underwent a 363% increase between 2000 and 2009 [DBC09]. In the last 10 years, neighboring fields of science with a longer history have influenced this growth of the MIR field. From the initial diversity of backgrounds and disciplines, not all did experience the same influence in the growth of the field. Looking back on the first 10 years of MIR shows a clear predominance of bottom-up methodologies issued from data-intensive disciplines such as Speech Processing, Text Retrieval and Computer Vision, as opposed to knowledge-based disciplines such as Musicology or (Music) Psychology. One possible reason for the relative stronger influence of data-intensive disciplines over knowledge-based ones is that the initial years of MIR co-occur with phenomena such as industrial applications of audio compression research and the explosive growth in the availability of data though the Internet (including audio files -mostly MP3s) [DBC09]. Further, following typical tasks from Speech Processing, Computer Vision and Text Retrieval, MIR research rapidly focused on a relatively small set of preferential tasks as local feature extraction, data modeling for comparison and classification, and efficient retrieval. In the following, we will review data processing methods employed in the three above-mentioned disciplines and relate their domains to the music domain to point out how MIR could benefit from further “Cross-Fertilisation” [Auc06, pp. 51-52] with these disciplines.

Speech Processing legacy

Speech Processing aims at extracting information from speech signals. This field has a long history and has been influential in a number of MIR developments, namely transcription, source recognition and source separation.

Music representation has been influenced by speech transcription and speaker recognition. It is commonplace to start any analysis of musical audio by the extraction of a set of local features, typical of speech transcription and speaker recognition, such as MFCCs computed on short-term Fourier transforms. In speech processing, these features make up the basic building blocks of machine learning algorithms that map patterns of features to individual speakers or likely sequences of words in multiple stages (i.e. short sequences of features mapped to phones, sequences of phones mapped to words and sequences of words mapped to sentences). A prevalent technique for mapping from one stage to the next being Hidden Markov Model (HMMs). Similar schemes have been adapted to music audio data and nowadays form the basis of music signal classification in genres [Auc06], tags or particular instruments.

Research in speech processing has also addressed the problem of separating out a single voice from a recording of many people speaking simultaneously (known as the "cocktail party" problem). A parallel problem when dealing with music data is isolating the components of a polyphonic music signal. Source separation is easier if there are at least as many sensors as sound sources (see [Mit04]). But in MIR, a typical research problem is the under-determined source separation of many sound sources in a stereo or mono recording. The most basic instantiation of the problem assumes that N source signals are linearly mixed into M < N channels, where the task is to infer the signals and their mixture coefficients from the mixed signal. To solve it, the space of solutions has to be restricted by making further assumptions, leading to different methods: Independent Component Analysis (ICA) assumes the sources to be independent and non-Gaussian, Sparse Component Analysis (SCA) assumes the sources to be sparse, and Non-negative Matrix Factorization (NMF) assumes the sources, coefficients and mixture to be nonnegative. Given that speech processing and content-based MIR both work in the audio domain, local features can be directly adopted – and in fact, MFCCs have been used in music similarity estimation from the very beginning [Foo97]. HMMs have also been employed for modeling sequences of audio features or symbolic music [FPW05]. Several attempts have been made to apply source separation techniques to music, adding domain-specific restrictions on the extracted sources to improve performance: [VK02] assume signals to be harmonic, [Vir07] assumes continuity in time, and [Bur08] incorporates instrument timbre models.

Text Retrieval legacy

Two tasks of Text Retrieval have had a great influence on MIR, namely document retrieval (in a given collection, find documents relevant to a textual query in the form of search terms or an example document) and document classification (assign a given document to at least one of a given set of classes, e.g., detect the topic of a news article or filter spam emails). Both problems require some abstract model for a document. The first system for document classification [Mar61] represented each document as a word count vector over a manually assembled vocabulary of "clue words", then applied a Naïve Bayes classifier to derive the document's topic, neither regarding the order nor co-occurrence of words within the document. Today, documents are still commonly represented as a word count vector – or *Bag of Words* (BoW) – for both classification and retrieval, but improvements over [Mar61] have been proposed on several levels, namely stemming, term weighting [SB88], topic modeling [DDL90] [Hof99] [BNJ03] [TJBB06] [SH09b], semantic hashing [HS11], word sense disambiguation [Nav09], and n-gram models (see [Seb02] for a review). Some of these techniques have been applied to find useful abstract representations of music pieces as well, but their use implies that a suitable equivalent to words can be defined for music. Some authors tried to apply vector quantization ("stemming") to frame-wise audio features ("words") to form a BoW model for similarity search [SWK08]. [RHG08] additionally employ TF/IDF term weighting of their so-called "audio-words". [HBC08] successfully applied HDP topic models for similarity estimation, albeit modeling topics as Gaussian distributions of MFCCs rather than multinomials over discrete words.

Computer Vision Systems legacy

Three typical Computer Vision problems have been particularly influential in MIR research, namely scene recognition (classifying images of scenery), multiple object detection (decomposing a complex image into a set of known entities and their locations) and image retrieval by example. Again, in Computer Vision, all these tasks require abstract representations of images or image parts to work with, and researchers have developed a wealth of image-specific local features and global descriptors (see [DJLW08], pp.17-24 for a review). A common framework has been inspired by Text Retrieval: [ZRZ02] regard images as documents composed of *keyblocks*, in analogy to text composed of keywords. Keyblocks are vector-quantized image patches extracted on a regular grid, forming a 2-dimensional array of "visual words", which can be turned into a *Bag of Visual Words* (BoVW) by building histograms. Several improvements have since been proposed, namely regarding visual words [vGGVS08] [CN11], Pooling [BPL10], Spatial pyramids [LSP06], Topic modeling [SRE05, FFP05], Generative image models [CLN10] [KH11] [RH10] [LGRN09], Learning invariances [HKW11], Semantic hashing [KH11]. As for Speech and Text processing, some of these techniques have been adapted to the processing of music audio features, typically MFCCs. Examples include [Abd02, pp.114-115] who employs sparse coding to short spectrogram excerpts of harpsichord music, yielding note detectors. [CEK05] use Haar-like feature extractors inspired from object detection to discriminate speech from music. [PKSW10] apply horizontal and vertical edge detectors to measure the amount of harmonic and percussive elements. [LLPN09] apply Convolutional RBMs for local feature extraction with some success in genre classification. [SO11] learn local iamge features for music similarity estimation. Additionally, as music pieces can be represented directly as images, by using e.g. images of spectrograms, several authors directly applied image processing techniques to music: [DSN01] extract features for genre classification with oriented Difference of Gaussian filters. Recent improvements on using image features for music classification can be found in [COKG11].

References

- [DBC09] J. Stephen Downie, Donald Byrd and Tim Crawford, "Ten Years of ISMIR: Reflections on Challenges and Opportunities", ISMIR 2009 http://ismir2009.ismir.net/proceedings/keynote1.pdf
- [Auc06] Jean-Julien Aucouturier. Dix Expériences sur la Modélisation du Timbre Polyphonique. PhD thesis, University of Paris 6, Paris, France, May 2006.
- [Mar61] M.E. Maron. Automatic indexing: an experimental inquiry. Journal of the Association for Computing Machinery, 8(3):404–417, 1961.
- [Seb02] Fabrizio Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47, 2002.
- [SB88] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5):513–523, 1988.
- [DDL90] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science, 41(6):391–407, 1990.
- [Hof99] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99), pages 50–57, 1999.
- [BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993–1022, 2003.
- [TJBB06] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet Processes. Journal of the American Statistical Association, 101(476):1566–1581, Dec 2006.

- [SH09b] Ruslan Salakhutdinov and Geoffrey Hinton. Replicated softmax: an undirected topic model. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, Advances in Neural Information Processing Systems 22 (NIPS 2009), pages 1607–1614. 2009.
- [HS11] Geoffrey Hinton and Ruslan Salakhutdinov. Discovering Binary Codes for Documents by Learning Deep Generative Models. Topics in Cognitive Science, 3(1):74–91, 2011.
- [Nav09] Roberto Navigli. Word Sense Disambiguation: A Survey. ACM Computing Surveys, 41:10:1–10:69, Feb 2009.
- [SWK08] Klaus Seyerlehner, Gerhard Widmer, and Peter Knees. Frame-level Audio Similarity – A Codebook Approach. In Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08), Espoo, Finland, 2008.
- [RHG08] Matthew Riley, Eric Heinen, and Joydeep Ghosh. A text retrieval approach to content-based audio retrieval. In Proceedings of the 7th International Society of Music Information Retrieval Conference (ISMIR 2008), pages 295–300, 2008.
- [HBC08] Matthew Hoffman, David M. Blei, and Perry R. Cook. Content-Based Musical Similarity Computation using the Hierarchical Dirichlet Process. In Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR 2008), pages 349–354, 2008.
- [DJLW08] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv., 40:5:1–5:60, May 2008.
- [ZRZ02] Lei Zhu, Al Bing Rao, and Aldong Zhang. Theory of Keyblock-based Image Retrieval. ACM Transactions on Information Systems, 20(2):224–257, Apr 2002.
- [vGGVS08] Jan van Gemert, Jan-Mark Geusebroek, Cor J. Veenman, and Arnold W. M. Smeulders. Kernel codebooks for scene categorization. In Proceedings of the 10th European Conference on Computer Vision (ECCV 2008), volume 5304 of Lecture Notes in Computer Science, pages 696–709. Springer, 2008.
- [CN11] A. Coates and A. Ng. The Importance of Encoding Versus Training with Sparse Coding and Vector Quantization. In Proceedings of the 28th International Conference on Machine Learning (ICML 2011), 2011.
- [BPL10] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pages 111–118, Haifa, Israel, 2010.
- [LSP06] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), pages 2169–2178, 2006.
- [SRE05] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. Discovering Objects and their Localization in Images. In Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05), volume 1, pages 370–377, 2005.
- [FFP05] Li Fei-Fei and Pietro Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 2, pages 524–531, 2005.
- [CLN10] Adam Coates, Honglak Lee, and Andrew Y. Ng. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2010.

- [KH11] Alex Krizhevsky and Geoffrey E. Hinton. Using Very Deep Autoencoders for Content-Based Image Retrieval. In Proceedings of the 19th European Symposium on Artificial Neural Networks (ESANN 2011), Bruges, Belgium, 2011.
- [RH10] M. Ranzato and G. Hinton. Modeling Pixel Means and Covariances Using Factorized Third-Order Boltzmann Machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10), pages 2551–2558, 2010.
- [LGRN09] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations. In Proceedings of the 26th International Conference on Machine Learning (ICML 2009), pages 609–616, Montreal, Quebec, Canada, 2009.
- [HKW11] Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. Transforming Auto-Encoders. In Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN 2011), volume 6791 of Lecture Notes in Computer Science, pages 44–51, Espoo, Finland, 2011. Springer.
- [DSN01] Hrishikesh Deshpande, Rohit Singh, and Unjung Nam. Classification of Music Signals in the Visual Domain. In Proceedings of the 4th International Conference on Digital Audio Effects (DAFx-01), Limerick, Ireland, 2001.
- [COKG11] Costa, Y. and Oliveira, L. and Koerich, A. and Gouyon, F. Music Genre Recognition Using Spectrograms. International Conference on Systems, Signals and Image Processing, 2011.
- [Abd02] Samer A. Abdallah. Towards Music Perception by Redundancy Reduction and Unsupervised Learning in Probabilistic Models. PhD thesis, King's College London, London, UK, 2002.
- [CEK05] Norman Casagrande, Douglas Eck, and Balázs Kégl. Frame-Level Speech/Music Discrimination using AdaBoost. In Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR 2005), pages 345–350, 2005.
- [PKSW10] T. Pohle, P. Knees, K. Seyerlehner, and G. Widmer. A High-Level Audio Feature For Music Retrieval and Sorting. In Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10), Graz, Austria, 2010.
- [LLPN09] Honglak Lee, Yan Largman, Peter Pham, and Andrew Y. Ng. Unsupervised Feature Learning for Audio Classification using Convolutional Deep Belief Networks. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, Advances in Neural Information Processing Systems 22 (NIPS 2009), pages 1096–1104. 2009.
- [SO11] Jan Schlüter and Christian Osendorfer. Music Similarity Estimation with the Mean-Covariance Restricted Boltzmann Machine. In Proceedings of the 10th International Conference on Machine Learning and Applications (ICMLA 2011), Honolulu, USA. 2011.
- [JR91] B.H. Juang and L.R. Rabiner. Hidden Markov Models for Speech Recognition. Technometrics, 33(3), 1991.
- [Fur86] S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on Acoustics, Speech and Signal Processing, 34(1):52–59, 1986.
- [Rob94] Tony Robinson. An Application of Recurrent Nets to Phone Probability Estimation. IEEE Transactions on Neural Networks, 5(2):298–305, 1994.
- [MDH09] Abdel-rahman Mohamed, George Dahl, and Geoffrey Hinton. Deep Belief Networks for Phone Recognition. In NIPS 22 Workshop on Deep Learning for Speech Recognition, 2009.

- [JH11] Navdeep Jaitly and Geoffrey Hinton. Learning a better Representation of Speech Sound Waves using Restricted Boltzmann Machines. In Proceedings of the 36th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), 2011.
- [DRMH10] George E. Dahl, Marc'Aurelio Ranzato, Abdel-rahman Mohamed, and Geoffrey E. Hinton. Phone recognition with the mean-covariance restricted Boltzmann machine. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, Advances in Neural Information Processing Systems 23 (NIPS 2010), pages 469–477. 2010.
- [Bei11] Homayoon Beigi. Speaker recognition. In Jucheng Yang, editor, Biometrics. InTech, 2011.
- [Mit04] Nikolaos Mitianoudis. Audio Source Separation using Independent Component Analysis. PhD thesis, Queen Mary, University of London, 2004.
- [Foo97] Jonathan T. Foote. Content-based retrieval of music and audio. In Jay C. C. Kuo, Shih F. Chang, and Venkat N. Gudivada, editors, Multimedia Storage and Archiving Systems II (Proceedings SPIE), volume 3229, pages 138–147, 1997.
- [FPW05] Arthur Flexer, Elias Pampalk, and Gerhard Widmer. Hidden markov models for spectral similarity of songs. In Proceedings of the 8th International Conference on Digital Audio Effects (DAFx-05), Madrid, Spain, 2005.
- [VK02] Tuomas Virtanen and Anssi Klapuri. Separation of Harmonic Sounds Using Linear Models for the Overtone Series. In Proceedings of the 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'02), volume II, pages 1757–1760, Orlando, FL, USA, 2002.
- [Vir07] Tuomas Virtanen. Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria. IEEE Transactions on Audio, Speech, and Language Processing, pages 1066–1074, 2007.
- [Bur08] Juan José Burred. From Sparse Models to Timbre Learning: New Methods for Musical Source Separation. PhD thesis, Technical University of Berlin, Berlin, Germany, Sept 2008.

2.4 Knowledge-driven methodologies

Our community has for a long time been focusing on a range of bottom-up approaches: starting with the kinds of data we use (see 2.1 and 2.2) to the types of algorithms we apply to it (see 2.3). This chapter is about transcending this focus and exploring other methodologies and fields of science which approach music in a more integral way. This includes musicology, psychology, sociology, neuroscience and human computer interaction. We ask ourselves what additional information about the process of music information research we can gain from these fields of science and how we can make the most of it. The focus is here on gaining domain knowledge from outside of MIR as opposed to borrowing methodologies or algorithms as discussed in 2.3.

2.4.1 State of the art

Musicology

Musicology, understood as the academic study of music [Harper-Scott & Samson 2009], is a main discipline in Music Information Research, music being our main object of study. For that reason, musicologists have taken an active role in the ISMIR community. For instance, the 2010 ISMIR Edition was hosted by a musicology department (http://ismir2010.ismir.net) and musicology has been considered as a key topic in the ISMIR call for papers (see, e.g. research areas related to computational musicology, computational ethnomusicology explicitly considered at http://ismir2012.ismir.net/authors/call-for-participation). Moreover, the conference on Interdisciplinary Musicology (CIM http://www.uni-graz.at/~parncutt/cim) has included papers on computational modeling in the program, and there is a special edition of this conference on the topic of "Technology" that is planned for 2014 http://www.sim.spk-berlin.de/cim14_919.html). There are also some relevant journals in this intersection (e.g. Journal of Mathematics and Music http://www.tandfonline.com/toc/tmam20) and the Special Issue of Computational Ethnomusicology at the Journal of New Music Research (http://www.tandfonline.com/toc/nnmr20). An overview on the relationship betweeen MIR and musicology is provided in the Musicology tutorial presented at ISMIR 2011 [Volk & Wiering, 2011]).

Although musicological studies in the ISMIR area have traditionally focused on the symbolic domain (section 2.1.1), recent developments in music transcription and feature extraction technologies from audio signals have opened new research paths on the intersection of musicology and signal processing. Key research topics in this area have been, among others, melodic similarity, key estimation and chord tracking. [Volk & Wiering 2011] contrasted (p. 46) musicological and MIR research in terms of, among others, data sources, repertoires and methodologies, and pointed out some opportunities for future research. MIR technologies can contribute with tools and data that are useful for musicological purposes, and Musicology can provide relevant research problems and use cases that can be addressed through MIR technologies. A mutual influence is starting to take place, although there is still a need for more collaboration between musicologists and technicians to create a truly interdisciplinary research area and contribute with truly music-rooted models and technologies. Only by this collaboration we can address the current gap between feature extractors and expert analyses and make significant contributions to existing application needs, e.g. version identification, plagiarisim detection [Cook & Sapp 2009], music recommendation, and to study how the relationship between people and music changes with the use of technology (e.g. "Musicology for the masses" project http://www.elec.qmul.ac.uk/digitalmusic/m4m/).

Psychology of Music

Music is created and experienced by humans, and the ultimate goal of MIR is to produce results that are helpful and interesting for humans. Therefore it is only natural to care about how humans perceive and create music. Music psychology tries to explain both musical behavior and musical experience with psychological methods. Its main instrument therefore is careful experimentation involving human subjects engaged in some kind of musical activity. Research areas span the whole spectrum from perception to musical interaction in large groups. Research questions concern the perception of sound or sound patterns, as well as perception of more musically meaningful concepts like harmony, pitch, rhythm, melody and tonality. The emotions associated with personal music experience are a part of music psychology, as are personal musical preferences and how they are influenced through peer groups and family, and musical behaviors from dancing to instrument playing to the most sophisticated interaction within whole orchestras.

Therefore music psychology should be able to provide valuable knowledge for MIR researchers in a whole range of sub-fields. Indeed there already is a certain exchange of knowledge between music psychology and MIR. Just to give a few examples, Carol L. Krumhansl, an eminent figure in music psychology, was an invited speaker at the Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010), in Utrecht, Netherlands (http://ismir2010.ismir.net/) talking about "Music and

Cognition: Links at Many Levels". Her monograph on "cognitive foundations of musical pitch" [Krumhansl 1990] is still seen as one of the standard texts on the subject. Gerhard Widmer, who has been an important contributor to MIR right from its start, will be a keynote speaker at the "12th International Conference on Music Perception and Cognition (ICMPC)" (http://icmpc-escom2012.web.auth.gr/), which is one of the most important conferences in the field of music psychology. And, last but not least, there are MIR researchers who contribute to both music psychology and MIR literature (e.g. Simon Dixon, http://www.eecs.qmul.ac.uk/~simond/).

Sociology of Music

Social psychology and the sociology of music focus on individuals as members of groups and on how groups and shared cultural codes influence music-related attitudes and activities. This point of view allows one to ask and answer important questions like: How do individuals and groups use music? How is the collective production of music made possible? How does music relate to broader social distinctions, especially class, race, and gender?

Although it is evident that such a sociology of music should be able to provide important insights not only for the field of MIR, many authors have suggested that research over recent decades has largely ignored the social functions of music at the expense of its cognitive and emotional functions (see e.g. [Hargreaves & North 1997]). [Hargreaves & North 1999] concluded that music serves three social functions: it is used by individuals to help manage their moods, self-identity [DeNora 2000], and interpersonal relationships. [North et al. 2000] supported this idea, showing that a sample of 13- to 14-year-olds listened to music to portray a social image to others, and to fulfill their emotional needs. Similarly, [Tarrant et al 2000] showed that American and English adolescents listened to music to satisfy both emotional and social needs, as well as for reasons of self-actualization. [Londsdale & North 2011] remarked that listening to music was "a social activity", which offered an opportunity for participants "to socialize with friends" (e.g., dancing, sharing live music). Even though music has a stronger social component for teenagers and young people than for seniors, its powers to strengthen social bonds and provide memory aids when brain functions decline are yet to be explored and exploited. How MIR can benefit from these and other results concerning the sociology of music is still a largely open question which opens up new and promising areas of research.

Neuroscience

All music psychological questions raised above could of course also be examined with neuroscientific methods. Instead of measuring the subject's behavior in music psychological experiments or directly asking subjects about their experiences concerning music it is possible to measure various signals from the human brain during such experiments. Possible signals range from electro-encephalography (EEG) to magneto-encephalography (MEG) or functional magnetic resonance imaging (fMRI). Each of the signals has their own characteristic strengths and weaknesses. E.g. EEG has a very good temporal but poor spatial resolution where fMRI is just the opposite. No matter what brain signals are being used, the fundamental question is always what parts of the brain contribute in what way to a subject's experience or creation of music. It is not immediately clear what MIR could gain from such a knowledge about brain structures involved in perception and production of music that could go beyond knowledge obtained from psychological experiments not utilizing neuroscientific methods. The biggest contribution might concern problems where humans have a hard time self-assessing their performance and experience. One example is the experience of emotions when listening to music. Neuroscientific methods might be able to provide a more quantitative and maybe more accurate picture than human self-assessment (see e.g. [Blood & Zatorre 2001], [Schmidt & Trainor 2001]). Differences in brain structure and function between skilled musicians and non-musicians is another well researched subject (see e.g. [Gaser & Schlaug 2003], [Krings et al. 1999]). The same holds for the study of the neuronal processes during performance of music where the sensorimotor interplay is at the center of interest (see [Zatorre et al. 2007] for a recent review).

Human Computer Interaction / Interfaces

The Association for Computing Machinery defines human-computer interaction (HCI) as "a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them."[1] HCI involves the study, planning, and design of the interaction between people (users) and computers. It is often regarded as the intersection of computer science, behavioral sciences, design and several other fields of study. Interaction between users and computers occurs at the interface, which includes both software and hardware. The basic and initial goal of HCI is to improve the interactions between users and computers by making computers more usable and receptive to the user's needs. In that sense, it is not surprising that for decades, HCI has mostly focused on making interaction more efficient, apparently not discovering until quite recently that not only efficiency is relevant, but perhaps also beauty and fun, and that the same perception or experience of interaction can have intrinsic values per se, independently of the outcomes it may produce [e.g. Norman 2004; McCarthy and Wright 2004].

Since, as we have already repeatedly stated, the human components of MIR are highly relevant, both from cultural, psychological or physiological perspectives, it comes as no surprise that MIR could strongly benefit from knowledge inherited from HCI and other related disciplines such as User Experience (UX), or Interface and Interaction Design studies. These methodologies could bring important benefits not only to the conception of MIR systems at earlier design stages, but also for the evaluation and subsequent iterative refinement of these systems. In that sense, whereas the evaluation of MIR systems conceived for providing univocally correct answers (e.g finding or identifying a known target song) seems quite straightforward, more open systems, with more open and thus more subjective outcomes, the inclusion of more subjective aspects such as the users' emotions, perceptions and internal states [e.g. Hekkert 2006], the effort done by users in order to accomplish an assigned action, or the context or environment within which the interaction occurs [e.g. Hassenzahl & Tractinsky 2006], could also be taken into account[2].

Furthermore, beyond the evaluation of User Experience, another MIR component that could directly benefit from HCI-related knowledge would be the application of SOA interface and interaction technologies in the creation of MIR systems and tools. This topic is covered in more detail in section 2.6.

[1] ACM SIGCHI Curricula for Human-Computer Interaction: http://old.sigchi.org/cdg/cdg2.html#2_1

[2] The current SOA in MIR **evaluation of research results** is covered in section 2.7

References

- [Cook & Saap 2009] N. Cook & C. Sapp. Purely coincidental? Joyce Hatto and Chopin's Mazurkas, online resource, http://www.charm.kcl.ac.uk/projects/p2_3_2.html#d3419e8182 .
- [Blood & Zatoore 2001] A.J. Blood, R.J. Zatorre. Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. Proceedings of the National Academy of Sciences of the United States of America 98 (20): 11818-11823, 2001.
- [DeNora 2000] T. DeNora, Music as a Technology of Self, chapter 3 of Music and Everyday Life, Cambridge: Cambridge University Press: pp. 46-74, 2000.
- [Gaser & Schlaug 2003] C. Gaser, G. Schlaug. Brain structures differ between musicians and non-musicians. The Journal of Neuroscience 23 (27): 9240-9245, 2003.
- [Hargreaves & North 1997] D.J. Hargreaves, A.C. North, A. C. The social psychology of music. Oxford: Oxford University Press, 1997.
- [Hargreaves & North 1999] D.J. Hargreaves, A.C. North. The functions of music in everyday life: Redefining the social in music psychology. Psychology of Music, 27, 71-83, 1999.
- [Harper-Scott & Samson 2009] . Harper-Scott, J. & Samson, J. (ed.) An Introduction to Music Studies., Cambridge and New York : Cambridge University Pres, 2009.
- [Hassenzahl and Tractinsky 2006] Hassenzahl, M., Tractinsky, N. (2006) "User experience – a research agenda". Behavior & Information Technology. Volume 5, Issue 2.
- [Hekkert, 2006] Hekkert, P (2006). "Design aesthetics: Principles of pleasure in product design". Psychology Science, 48(2), 157-172 (2006).
- [Krings et al 1999] T. Krings, R. Topper, H. Foltys, S. Erberich, R. Sparing, K. Willmes, A. Thron. Cortical activation patterns during complex motor tasks in piano players and control subjects. A functional magnetic resonance imaging study. Neuroscience Letters 278 (3): 189-193, 1999.
- [Krumhansl 1990] C.L. Krumhansl. Cognitive Foundations of Musical Pitch. Oxford University Press, USA, 1990.
- [Lonsdale & North 2011] A.J. Londsdale, A.C. North, Why do we listen to music? A uses and gratifications analysis. British Journal of Psychology, 102, 108-134, 2011.
- [McCarthy & Wright 2004] McCarthy, J. and Wright, P. (2004) "Technology as experience", Interactions, Vol. 11, No. 5, 42-43.
- [Norman 2004] Norman, D. (2004). Emotional Design: Why We Love (Or Hate) Everyday Things, Basic Books.
- [North et al 2000] A.C. North, D.J. Hargreaves, S.A. O'Neill. The importance of music to adolescents. British Journal of Educational Psychology, 70, 255-272, 2000.
- [Schmidt & Trainor 2001] L.A. Schmidt, L.J. Trainor. Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions. Cognition and Emotion 15 (4): 487-500, 2001.
- [Tarrant et al 2000] M. Tarrant, A.C. North, D.J. Hargreaves, English and American adolescents' reasons for listening to music. Psychology of Music, 28, 166-173, 2000.
- [Volk & Wiering 2011] A. Volk, F. Wiering. Musicology. Tutorial, ISMIR 2011.
- [Zatorre et al 2007] R.J. Zatorre, J.L. Chen, V.B. Penhune. When the brain plays music. Auditory-motor interactions in music perception and production. Nature Reviews Neuroscience, 8, 547-558, 2007.

2.5 Music content analysis

Audio Content Analysis (ACA) denotes the research domain aiming at extracting music content using algorithms applied to the audio signal. One of its goals is to estimate the score of a music track (melody, harmony, rhythm, beat and downbeat positions, overall structure) from the audio signal. ACA has been a major field of research in the MIR community over the past decade. But how can algorithm performance be further improved in this field?

2.5.1 State of the art

Mostly, ACA algorithms can be divided into two groups with different aims, different practices and which usually involve different research centers.

The machine-learning group (ML-group) aims at extracting subjective or application-oriented information (such as genre, mood, user tags or similarity). The current practice in this domain is to derive knowledge automatically from so-called low-level features (MFCC, ZCR, Spectral Flatness) of data with known class labels, using machine-learning rules or models (SVM, AdaBoost, RandomForest). The obtained rules/ models are then applied to unknown data to propagate knowledge (to infer genre, mood or user tags).

The aim of the second group, the signal-processing group (SP-group), is to estimate the parameters related to written music. Since the MIR community is until now largely made up of Western researchers, written music refers to the music notation system originated from European classical music, consisting of notes with an associated position and duration inside a bar, in the context of a meter (hence providing beat positions inside the bars), a clef (indicating the octaves assigned to the notes), a key signature (series of sharps or flats) organized by instrument (or hands) into parallel staffs, and finally organized in a large structure corresponding to musical movements. An extension of common notation summarizes groups of simultaneously occurring notes using chord symbols. ACA aims at retrieving this music notation from the observation of an audio music track (realization of a generative music process). In this, all the notes of the parallel staffs occur simultaneously in an interpretative process. These two points (simultaneous occurrence and interpretation) infer that ACA for music transcription cannot be solved using methods from the ML-group. It will necessitate the development of elaborated signal-processing algorithms (SP-group) and developing manual mapping strategies. Since the audio signal represents a realization of the music notation it exhibits variations in terms of interpretation (not all the notes are played, pitches vary over time, musicians modify timing). ACA algorithms estimate pitches with associated starting and ending times which are then mapped to the [note-height, clef, key, note position and duration] system. All this makes music transcription a difficult problem to solve. For this reason, the number of research related to this field is largely below the one related to the ML-group. Moreover, until recently, from an application point-of-view, the market place was considered limited (to users with musical training). Today, with the success of applications such as Melodyne (multi-pitch), Garage-Band, the need for search using Query-by-Humming (dominant melody extraction), mobile applications such as Tonara (iPad) or online applications such as Song2See (web-based), information related to music transcription is now reaching everyday people.
For the estimation of music transcription two major trends can be distinguished.

A. Non-informed estimation (estimation-from-scratch)

These approaches attempt to estimate the various music score concepts from scratch (without any information such as score or chord-tabs). In this category, approaches have been proposed for estimating the various pitches, the key, the sequence of chords, the beat and downbeat positions and the global structure.

Multi-pitch estimation is probably the most challenging task since it involves being able to separate the various pitches occurring simultaneously and estimating the number of sources playing at any time. According to [Yeh et al., 2010], most multi-pitch algorithms follow three main principles closely related to mechanisms of the auditory system: harmonicity, spectral smoothness, and synchronous amplitude evolution within a given source. From these principles a number of approaches are derived: solving the problem using a global optimization scheme such as NMF [Vincent et al., 2008], harmonic temporal structured clustering [Kameoka et al., 2007], iterative optimization [Klapuri, 2008] or a probabilistic framework [Ryynanen and Klapuri, 2008; Emiya et al., 2008]. Considering the fact that the performance obtained in the past years in the related MIREX task (~69% note-accuracy) remains almost constant, it seems that a glass ceiling has been reached in this domain and that new approaches should be studied.

Key and chord estimation are two closely related topics. They both aim at assigning a label among a dictionary (a fixed set of 24 tonalities, or the various triads with possible extensions) to a segment of time. Given that the estimation of key and chords from estimated multi-pitch data is still unreliable (see [Papadopoulos, 2010]) algorithms rely for the most part on the extraction of Chroma or Harmonic Pitch Class Profiles [Gomez, 2006] possibly including harmonic/pitch-enhancement or spectrum whitening. Then, a model (either resulting from perceptual experiments, trained using data or inspired by music theory) is used to map the observations to the labels. In this domain, the modeling of dependencies (with HMM or Bayesian networks) between the various musical parameters is a common practice: dependencies between chords and key [Pauwels and Martens, 2010; Rocher et al., 2010], between successive chords, between chord, metrical position and bass-note [Mauch and Dixon, 2010], or between chord and downbeat [Papadopoulos and Peeters, 2010]. Key and chord estimation is the research topic that relies the most on music theory.

While music scores define the temporal grid at the bar/measure level; most research focuses on the beat level (named tactus). Only recent research try to directly estimate the bar/measure positions (named downbeat). In this field, methods can be roughly subdivided into a) audio-to-symbolic or onset-based methods and b) energy-variation-based methods [Scheirer, 1998]. The periodicities can then be used to infer the tempo directly or to infer the whole metrical structure (tatum, tactus, measure, systematic time deviations such as swing factor [Laroche, 2003]) through probabilistic or multi-agent models. Other sorts of front-ends have also been used to provide higher-level context information (chroma-variation, spectral balance [Goto, 2001], [Klapuri et al., 2006] [Peeters and Papadopoulos, 2011]). Recent methods propose the use of neural networks to learn the specific characteristics of beat positions with very good results [Boeck and Schedl, 2011]. This field is still very active and creative. Given the importance of correct estimation of the musical time-grid provided by beat and downbeat information, this field will remain active for some time. A good overview can be found in [Gouyon and Dixon, 2005].

Research on the estimation of Music Structure from audio started at the end of the ‘90s with the work of Foote [Foote, 1999] (co-occurrence matrix) and Logan [Logan and Chu, 2000]. By ”structure” the various works mean detection of homogeneous parts (state approach) or repetitions of sequences of events, possibly including transpositions or time-stretching (sequence approaches [Peeters, 2004]). Both methods share the use of low-level features such as MFCC or Chroma/PCP as front-end. State methods are usually based on time-segmentation and various clustering or HMM techniques [Levy and Sandler, 2008]. Sequence approaches usually first detect repetitions in a self-similarity matrix and then infer the structure from the detected repetitions using heuristics or fitness approaches [Paulus and Klapuri, 2009]. Relationships between the structure and the various other content-elements can be found in the use of beat-synchronous features, and the use of structure to reinforce tempo or chord estimation [Mauch et al., 2009]. Good overviews of this topic can be found in [Dannenberg and Goto, 2009] and [Paulus et al., 2010].

B. Informed estimation (alignment and followers)

These approaches use previous information (such as given by a score, a MIDI file or a text-transcription) and align it to an audio file hence providing inherently its estimation. This method is currently applied to two fields for which estimation-from-scratch remains very complex: scores and lyrics.

Score alignment and score following are two closely related topics in the sense that the latter is the real-time version of the first. They both consist in finding a time-synchronization between a symbolic representation and an audio signal. Historically, score following was developed first with the goal of allowing interactions between a computer and a musician ([Dannenberg, 1984], [Vercoe, 1984]) using MIDI or fingering information and not audio because of CPU limits. Works were later extended by Puckette [Puckette, 1990] to take into account pitch estimation from audio and deal with polyphonic data. Given the imperfect nature of observations, [Grubb and Dannenberg, 1997] introduced statistical approaches. Since 1999, Hidden Markov Model/ Viterbi seems to have been chosen as the main model to represent time dependency [Cano et al., 1999], [Raphael, 1999]. The choice of Viterbi decoding, which is also used in dynamic time warping (DTW) algorithms, is the common point between Alignment and Followers [Orio and Schwarz, 2001]. Since then, the focuses of the two fields have been different. Alignment focuses on solving computational issues related to DTW [Müller et al., 2006], and Follower on anticipations (using tempo or recurrence information [Cont, 2008]). While being the privilege of a limited number of people, today score following is now accessible by the large-audience through recent applications such as Tonara (iPad) or Songs2See (web-based).

Automatic transcription of the lyrics of a music track is another complex task. It involves first locating the signal of the singer in the mixed audio track, and then recognizing the lyrics conveyed by this signal (large differences between the characteristics of the singing voice and speech make standard speech transcription systems unsuitable for the singing voice). Works on alignment started with the isolated singing voice [Loscos et al., 1999] and were later extended to the singing voice mixed with other sources. Usually systems first attempt to isolate the singing voice (e.g. using the PreFest dominant melody detection algorithm [Fujihara et al., 2011]), then estimate a Voice Activity Criterion and then decode the phoneme sequence using a modified HMM topology (filler model in [Fujihara et al., 2011]), adapting the speech phoneme model to singing. Other systems also exploit the temporal relationships between the text of the lyrics and the music. For example, the system Lyrically [Wang et al., 2004] uses the specific assumption that lyrics are organized in paragraphs as the music is organized in segments. The central segment being the chorus will serve as anchor-point. Measure positions are used as the anchor-point for lines. [Mauch et al., 2012] use the relationship between lyrics and chords to strengthen lyrics synchronization by chord estimation.

References

- [Boeck and Schedl, 2011] Boeck, S. and Schedl, M. (2011). Enhanced beat tracking with context-aware neural networks. In Proc. of DAFx, Paris, France.
- [Cano et al., 1999] Cano, P., Loscos, A., and Bonada, J. (1999). Score-performance matching using hmms. In Proc. of ICMC, Bejing, China.
- [Cont, 2008] Cont, A. (2008). Antescofo: Anticipatory synchronization and control of interactive parameters in computer music. international computer music conference. In Proc. of ICMC, Belfast, Ireland.
- [Dannenberg, 1984] Dannenberg, R. (1984). An on-line algorithm for real-time accompaniment. In Proc. of ICMC, pages 193–198. Computer Music Association.
- [Dannenberg and Goto, 2009] Dannenberg, R. and Goto, M. (2009). Music structure analysis from acoustic signal. In Handbook of Signal Processing in Acoustics Vol. 1, pages 305–331. Springer Verlag.
- [Emiya et al., 2008] Emiya, V., Badeau, R., David, B., et al. (2008). Automatic transcription of piano music based on hmm tracking of jointly-estimated pitches. In Proc. of EUSIPCO, Lausanne Switzerland.
- [Foote, 1999] Foote, J. (1999). Visualizing music and audio using self-similarity. In Proc. of ACM Int. Conf. on Multimedia, pages 77–80, Orlando, Florida, USA.
- [Fujihara et al., 2011] Fujihara, H., Goto, M., Ogata, J., and Okuno, H. (2011). Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics. IEEE Journal of Selected Topics in Signal Processing, 5(6):1252–1261.
- [Gomez, 2006] Gomez, E. (2006). Tonal description of polyphonic audio for music content processing. INFORMS Journal on Computing, Special Cluster on Computation in Music, 18(3).
- [Goto, 2001] Goto, M. (2001). An audio-based real-time beat tracking system for music with or without drum-sounds. Journal of New Music Research, 30(2):159–171.
- [Gouyon and Dixon, 2005] Gouyon, F. and Dixon, S. (2005). A review of rhythm description systems. Computer Music Journal, 29(1):34–54.
- [Grubb and Dannenberg, 1997] Grubb, L. and Dannenberg, R. (1997). A stochastic method of tracking a vocal performer. In Proc. of the ICMC, pages 301–308, Thessaloniki, Greece.
- [Kameoka et al., 2007] Kameoka, H., Nishimoto, T., and Sagayama, S. (2007). A multipitch analyzer based on harmonic temporal structured clustering. IEEE Transactions on Audio, Speech and Language Processing, 15(3):982–994.
- [Klapuri, 2008] Klapuri, A. (2008). Multipitch analysis of polyphonic music and speech signals using an auditory model. IEEE Transactions on Audio, Speech and Language Processing, 16(2): 255–266.
- [Klapuri et al., 2006] Klapuri, A., Eronen, A., and Astola, J. (2006). Analysis of the meter of acoustic musical signals. IEEE Transactions on Audio, Speech and Language Processing, 14(1): 342–355.
- [Laroche, 2003] Laroche, J. (2003). Efficient tempo and beat tracking in audio recordings. J. Audio Eng. Soc., 51(4):226–233.
- [Levy and Sandler, 2008] Levy, M. and Sandler, M. (2008). Structural segmentation of musical audio by constrained clustering. IEEE Transactions on Audio, Speech and Language Processing, 16(2):318–326.

- [Logan and Chu, 2000] Logan, B. and Chu, S. (2000). Music summarization using key phrases. In Proc. of IEEE ICASSP, volume II, pages 749–752, Istanbul, Turkey.
- [Loscos et al., 1999] Loscos, A., Cano, P., and Bonada, J. (1999). Low-delay singing voice alignment to text. In Proc. of ICMC, page 23, Bejing, China.
- [Mauch and Dixon, 2010] Mauch, M. and Dixon, S. (2010). Simultaneous estimation of chords and musical context from audio. IEEE Transactions on Audio, Speech, and Language Processing, 18(6):1280 – 1289.
- [Mauch et al., 2012] Mauch, M., Fujihara, H., and Goto, M. (2012). Integrating additional chord information into hmm-based lyrics-to-audio alignment. IEEE Transactions on Audio, Speech and Language Processing, (99):1–1.
- [Mauch et al., 2009] Mauch, M., Noland, K., and Dixon, S. (2009). Using musical structure to enhance automatic chord transcription. In Proc. of ISMIR, Kobe, Japan.
- [Mueller et al., 2006] Mueller, M., Mattes, H., and Kurth, F. (2006). An efficient multiscale approach to audio synchronization. In Proc. ISMIR, pages 192–197, Victoria, Canada.
- [Orio and Schwarz, 2001] Orio, N. and Schwarz, D. (2001). Alignment of monophonic and polyphonic music to a score. In Proc. of ICMC, Havana, Cuba.
- [Papadopoulos, 2010] Papadopoulos, H. (2010). Joint Estimation of Musical Content Information. Phd thesis, University Paris VI.
- [Papadopoulos and Peeters, 2010] Papadopoulos, H. and Peeters, G. (2010). Joint estimation of chords and downbeats from an audio signal. IEEE Transactions on Audio, Speech and Language Processing, 19(1):138 – 152.
- [Paulus and Klapuri, 2009] Paulus, J. and Klapuri, A. (2009). Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. IEEE Transactions on Audio, Speech and Language Processing, 17(6):1159–1170.
- [Paulus et al., 2010] Paulus, J., Muller, M., and Klapuri, A. (2010). Audio-based music structure analysis. In Proc. of ISMIR, Utrecht, The Netherlands.
- [Pauwels and Martens, 2010] Pauwels, J. and Martens, J.-P. (2010). Integrating musicological knowledge into a probabilistic system for chord and key extraction. In Proc. AES 128th Conv, London, UK.
- [Peeters, 2004] Peeters, G. (2004). Deriving Musical Structures from Signal Analysis for Music Audio Summary Generation: Sequence and State Approach, pages 142–165. Lecture Notes in Computer Science. Springer-Verlag Berlin Heidelberg 2004.
- [Peeters and Papadopoulos, 2011] Peeters, G. and Papadopoulos, H. (2011). Simultaneous beat and downbeat-tracking using a probabilistic framework: theory and large-scale evaluation. IEEE Transactions on Audio, Speech and Language Processing, 19(6):1754–1769.
- [Puckette, 1990] Puckette, M. (1990). Explode: A user interface for sequencing and score following. In Proc. of ICMC, pages 259–261.
- [Raphael, 1999] Raphael, C. (1999). Automatic segmentation of acoustic musical signals using hidden markov models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(4): 360–370.
- [Rocher et al., 2010] Rocher, T., Robine, M., Hanna, P., Oudre, L., et al. (2010). Concurrent estimation of chords and keys from audio. In Proc. of ISMIR, Ultrecht, The Nederlands.
- [Ryynanen and Klapuri, 2008] Ryynanen, M. and Klapuri, A. (2008). Automatic transcription of melody, bass line, and chords in polyphonic music. Computer Music Journal, 32(3):72–86.

- [Scheirer, 1998] Scheirer, E. (1998). Tempo and beat analysis of acoustic musical signals. J. Acoust. Soc. Am., 103(1):588–601.
- [Vercoe, 1984] Vercoe, B. (1984). The synthetic performer in the context of live performance. In Proc. ICMC, pages 199–200.
- [Vincent et al., 2008] Vincent, E., Berlin, N., and Badeau, R. (2008). Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. In Proc. of IEEE ICASSP, pages 109–112. IEEE.
- [Wang et al., 2004] Wang, Y., Kan, M., Nwe, T., Shenoy, A., and Yin, J. (2004). Lyrically: automatic synchronization of acoustic musical signals and textual lyrics. In Proceedings of the 12th annual ACM international conference on Multimedia, pages 212–219. ACM.
- [Yeh et al., 2010] Yeh, C., Roebel, A., and Rodet, X. (2010). Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. In IEEE Transactions on Audio, Speech and Language Processing, volume 18, page 6.

2.6 Interface and interaction aspects

In the last decade, Human Computer Interaction (HCI) research has witnessed a change in focus from conventional ways to control and communicate with computers (keyboards, joystick, mice, knobs, levers, buttons, etc.) to more natural non-conventional devices such as gloves, speech recognition, eye trackers, cameras, or tangible user interfaces. With the advent of the Internet and the ubiquity of personal computers in the nineties, the graphical user interface (GUI) emerged as the pervasive interface that both users and designers had to deal with. Yet, at the same time, albeit lesser known at that time, computing started to progressively move beyond the desktop into new physical and social contexts as a result of both technological advances and a desire to surpass the WIMP (window, icon, menu, pointing device) limitations. Nowadays, when words such as "multi-touch" and gestures like the "two-finger pinch zoom" are part of a user's daily life, novel areas such as "tangible interaction" seem to have finally entered the mainstream. Despite this, if we omit the ongoing research explicitly focused towards real-time musical performance, which typically falls under the New Interfaces for Musical Expression (NIME[3]) discipline, not much research has been yet devoted for applying novel interface and interaction concepts to the field of M.I.R.

2.6.1 State of the art

Interfaces to music collections

Over the past 10 years a few projects from the MIR community have witnessed a marked evolution in the development of interfaces for music search and discovery. In the field of visualization, there is an extensive bibliography on the representation of auditory data. In the particular case of the visual organization of musical data, solutions often consist in extracting feature descriptors from data files, and creating a multidimensional feature space that will be projected into a 2D surface, using dimensionality reduction techniques. A very well known example of this method is the work Islands of Music by [Pampalk 2003], which uses a landscape metaphor to present a large collection of musical files. In this work a Self Organizing Map (SOM) [Kohonen 2001] is used for creating an artificial map in which the

[3] www.nime.org

accumulation of songs is presented as the elevation of the terrain over the sea. The islands created as a result of this process roughly correspond to musical genres. A later attempt to combine different visualizations on a single map was also created by [Pampalk et al. 2004]. By using different parameters to organize the SOM, they created several views of the collection, later interpolating the different solutions for creating a smooth combination of situations with which to explore new information.

Beyond 2D views, nepTune, an interactively explorable 3D version of Islands of Music supporting spatialized sound playback [Knees et al. 2007], and the Globe of Music, which places a collection on spherical surface to avoid any edges or discontinuities [Leitich and Topf 2007], are worth being mentioned. More recently, MusicGalaxy [Stober and Nürnberger 2008; 2011] implements an adaptive zoomable interface for exploration that makes use of a complex non-linear multi-focal zoom lens and introduces the concept of facet distances representing different aspects of music similarity. In the aforementioned examples, a topological metaphor is taken in advantage to enable users exploring big collections of data. A different and original visualization approach is chosen in Musicream [Goto and Goto 2005], an interesting example of exploratory search in music databases, using the search by example paradigm. In Musicream, songs are represented using coloured circles, which fall down from the top of the screen. When selected, these songs show their title on their center, and they can be later used to 'fish' similar ones. A commercial product that makes use of MIR technology is the Bang & Olufsen BeoSound 5, a high-end Hi-Fi system interface component that automatically continues to play similar sounding music without requiring user interaction[4], and which uses audio similarity algorithms underlying this "more-of-the-same"-component (MOTS) that have been developed at the JKU in cooperation with OFAI.

HCI, Tangible and Tabletop Interaction

In the last decade, Human Computer Interaction (HCI) research has witnessed a change in focus from conventional ways to control and communicate with computers (keyboards, joystick, mice, knobs, levers, buttons, etc.) to more natural non-conventional devices such as gloves, speech recognition, eye trackers, cameras, or tangible user interfaces. Computing started to progressively move beyond the desktop into new physical and social contexts as a result of both technological advances and a desire to surpass the WIMP (window, icon, menu, pointing device) limitations.

Tangible User Interfaces (TUI), which combine control and representation in a single physical device [Ullmer and Ishii 2001], constitute one of this novel approaches. Whereas in direct manipulation with GUI, users interact with digital information by selecting graphic representations (icons, windows, etc.) with pointing devices, tangible interaction emphasizes tangibility and materiality, physical embodiment of data, bodily interaction and the embedding of systems in real spaces and contexts. Professor Ishii at the MIT MediaLab coined the term Tangible User Interface in 1997 [Ishii and Ullmer 1997], although several related research and implementations predate this concept. Ishii envisioned TUIs as interfaces meant to augment the real physical world by coupling digital information to everyday physical objects and environments, literally allowing users to grasp data with their hands, thus fusing the representation and control of digital data and operations with physical artefacts.

[4] http://www.cp.jku.at/people/widmer/BeoSound_OFAI_Backgrounder_official.pdf

Tabletop Interaction constitutes a special domain in Tangible Interaction. We can think of a tabletop interface as a horizontal surface meant to be touched and/or manipulated through objects on it. Typically, this type of interface allows more than one input event to enter the system at the same time; instead of having one mouse and one keyboard restricting the user's input to an ordered sequence of events (click, click, double click, etc.), in interactive tables, any action is possible at any time and position, by one or by several simultaneous users. This feature leads us to what arguably constitutes the most commercially successful capability of horizontal surfaces: multi-touch interaction. The other implicit capacity of table-shaped interfaces is the ability to literally support physical items on them. Users can interact with objects with volume, shape and weight, and when the tracking system is able to identify these objects, and track their position and orientation, the potential bandwidth and richness of the interaction goes far beyond the simple idea of multi-touch. Interacting with the fingers still belongs to the idea of pointing devices, while interacting with physical objects can take us much farther. Such objects can represent abstract concepts or real entities; they can relate to other objects on the surface; they can be moved and turned around on the table surface, and all these spatial changes can affect their internal properties and their relationships with neighbouring objects. The availability of open-source, cross-platform computer vision frameworks that allows the tracking of fiducial markers and combined multi-touch finger tracking, such as reacTIVision, which was developed for the Reactable project by one of the members of this consortium [Bencina et al. 2005], and which is nowadays widely used among the tabletop developers community (both academic and industrial), has spread the development of tabletop applications mainly for education and for creativity [e.g. Khandelwal and Mazalek 2007; Gallardo et al. 2008].

Tabletop Interfaces for M.I.R. applications

There is a growing interest in applying Tabletop Interfaces to the music domain. From the Audiopad [Patten et al. 2002] to the Reactable [Jordà et al. 2007], music performance and creation has arguably become the most popular and successful application field in the entire lifetime of this interaction paradigm. In this sense, and although less prolific than the applications strictly conceived for musical performance, some interesting works have also been developed bridging tabletop interaction with MIR, specially oriented for interacting with large music collections. Musictable [Stavness et al. 2005], takes a visualization approach similar to the one chosen in Pampalk's Islands of Music, for creating a two dimensional map that, when projected on a table, is used to make collaborative decisions to generate playlists. Another adaptation into the tabletop domain is the work from [Hitchner et al. 2007], which uses a SOM to build the map and also creates a low-resolution mosaic that is shown to the user. The users can redistribute the songs on this mosaic and thus changing the whole distribution of SOM according to the user's desires. The MTG's SongExplorer [Julià and Jordà 2009] addresses the problem of finding new interesting songs on large music databases, from an interaction design perspective.

Using high-level descriptors of musical songs, SongExplorer addresses N-Dimensional navigation in a 2D plane by creating a coherent 2D map based on similarity, in which neighbouring songs tend to be more similar. All songs are represented as throbbing circles that highlight their relevant high-level properties, and the resulting music map is browseable and zoomable by the users who can use their fingers as well as specially designed tangible pucks, for helping them to find interesting music, independently of their previous knowledge of the collection. Tests comparing the system with a conventional GUI interface controlling the same music collection, showed that the tabletop implementation was a much more efficient tool for discovering new, valuable music to the users.

In short, tabletop interfaces have proven their suitability and potential for real-time and complex musical interaction. This is given by the specific affordances of this type of interfaces: support of collaboration and sharing of control; continuous, real-time interaction with multidimensional data; and support of complex, expressive and explorative interaction [Jordà 2008]. It is expected that these type of interfaces, together with the more ubiquitous and easily available individual multi-touch devices, such as tablets and smart-phones, can bring novel approaches to the field of MIR, not only for music browsing but also specially in more creative aspects related to MIR music creation and performance (cfr. section 4.2).

References

- Bencina, R., Kaltenbrunner, M. and Jordà, S. (2005) "Improved Topological Fiducial Tracking in the reacTIVision System", Proceedings of the IEEE International Workshop on Projector-Camera Systems.
- Gallardo, D., Julià, C. F., and Jordà, S. (2008). TurTan: A tangible programming language for creative exploration. 2008 3rd IEEE International Workshop on Horizontal Interactive Human Computer Systems (pp. 89-92). IEEE.
- Goto M. and Goto. T. (2005). Musicream: New music play- back interface for streaming, sticking, sorting, and re- calling musical pieces. In ISMIR 2005: Proceedings of the 6th International Conference on Music Information Retrieval, 2005.
- Hitchner, S., Murdoch, J. and Tzanetakis G. (2007) Music browsing using a tabletop display. Conference on Music Information Retrieval ISMIR 2007.
- Ishii, H., & Ullmer, B. (1997). Tangible bits: towards seamless interfaces between people, bits and atoms. Conference on Human Factors in Computing Systems, 234.
- Jordà, S., Geiger, G., Alonso, M., & Kaltenbrunner, M. (2007). The reacTable: exploring the synergy between live music performance and tabletop tangible interfaces. Proceedings of the 1st international conference on Tangible and embedded interaction (pp. 139–146). ACM.
- Jordà, S. (2008). "On Stage: the Reactable and other Musical Tangibles go Real". International Journal of Arts and Technology, 1-3/4: 268-287.
- Julià, C. F., and Jordà, S. (2009). Songexplorer: A tabletop application for exploring large collections of songs. ISMIR 2009.
- Khandelwal, M., and Mazalek, A. (2007). Teaching table: a tangible mentor for pre-k math education. Proceedings of the 1st international conference on Tangible and embedded interaction (pp. 191–194). ACM.
- Knees, P., Pohle, T., Schedl, M., and Widmer, G. (2007). "Exploring Music Collections in Virtual Landscapes". IEEE MultiMedia, vol. 14, no. 3, pp.46–54.
- Kohonen. T. (2001) Self-Organizing Maps. Springer, 2001.
- Leitich, S. and Topf. M. (2007). Globe of Music: Music Library Visualization Using GEOSOM. In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007).
- Pampalk. E. (2003). Islands of Music Analysis, Organization, and Visualization of Music Archives. Journal of the Austrian Soc. for Artificial Intelligence, 22(4):20–23, 2003.
- Pampalk, E., Hlavac, P., and Herrera, P. (2004) "Hierarchical Organization and Visualization of Drum Sample Libraries". In the Proceedings of the 7th International Conference on Digital Audio Effects (DAFx'04), pp 378-383, Naples, Italy.
- Patten, J., Recht, B. and Ishii, H. (2002). Audiopad: A tag-based interface for musical performance. In Proceedings of the 2002 conference on New interfaces for musical expression, pages 1–6.
- Stavness, J. Gluck, L. Vilhan, and S. Fels. (2005) The MUSICtable: A Map-Based Ubiquitous System for Social Interaction with a Digital Music Collection. Lecture Notes In Computer Science, 3711:291.

- Stober, S. and Nürnberger, A. (2008) “Towards user-adaptive structuring and organization of music collections”. In Proceedings of the 6th international workshop on Adaptive Multimedia Retrieval (AMR’08).
- Stober, S. and Nürnberger, A. (2011). Exploring Music Contents. Lecture Notes in Computer Science, 2011, Volume 6684/2011, 273-302, DOI: 10.1007/978-3-642-23126-1_18.
- Ullmer B. and Ishii, H. (2001) “Emerging Frameworks for Tangible User Interfaces”, In Human-Computer Interaction in the New Millenium, Ed. John M. Carroll, Addison-Wesley, 579-601.

2.7 Evaluation of research results

In many scientific disciplines, the creation of knowledge and advancement of the field rely on the possibility for independent researchers to build upon previous research, which in turn depends on the proper understanding of current state-of-the-art of the field. Availability of scientific literature partly facilitates this understanding, but there are far more crucial factors to this such as: (i) reproducibility of research (i.e. where researchers provide entire environments necessary to reproduce results: the data, the computer code, etc), and (ii) systematic, public and large-scale evaluations of algorithms and systems. Indeed, in many scientific disciplines dealing with data, improvements over the long term often rely on such evaluations.

2.7.1 State of the art

Many experimental disciplines have witnessed significant improvements over the long term thanks to community-wide efforts in systematic evaluations. This is the case for instance of (text-based) Information Retrieval with the TREC initiative (Text REtrieval Conference see http://trec.nist.gov) and the CLEF initiative (Cross-Language Evaluation Forum, http://www.clef-initiative.eu/), Speech Recognition [1], Machine Learning [2], and Video and Multimedia Retrieval with e.g. the TRECVID (http://www-nlpir.nist.gov/projects/trecvid/) and VideoCLEF initiatives (the latter later generalised to the "MediaEval Benchmarking Initiative for Multimedia Evaluation", http://multimediaeval.org/).

Although evaluation “per se” has not been a traditional focus of pioneering computer music conferences (such as the ICMC) and journals (e.g. Computer Music Journal), recent attention has been given to the topic. In 1992, the visionary Marvin Minsky declared: “the most critical thing, in both music research and general AI research, is to learn how to build a common music database” [3], but this was not until a series of encounters, workshops and special sessions organised between 1999 and 2003 by researchers from the newly-born community of Music Information Retrieval that the necessity of conducting rigorous and comprehensive evaluations was recognised [4].

The first public international evaluation benchmark took place at the ISMIR Conference 2004 [5], where the objective was to compare state-of-the-art audio algorithms and systems relevant for some tasks of music content description. This effort has then been systematized and continued via the yearly Music Information Retrieval Evaluation eXchange (MIREX). MIREXes have widened the scope of the competitions and now cover a broad range of tasks, including symbolic data description and retrieval [6].

The number of evaluation endeavors issued from different communities (e.g. Signal Processing, Data Mining, Information Retrieval), yet relevant to Music Information Research have recently increased significantly. For instance, the Signal Separation Evaluation Campaign (SiSEC) was started in 2008 (http://sisec.wiki.irisa.fr/), and deals with aspect of source separation in signals of different natures (music, audio, biomedical, etc.). It appears to run now as an annual event. A Data Mining contest was organised at the 19th International Symposium on Methodologies for Intelligent Systems (ISMIS) with two tracks relevant to MIR research (Tunedit): Music Genre recognition and Music Instruments recognition (http://tunedit.org/challenge/music-retrieval). The CLEF initiative (an IR evaluation forum) extended its scope to MIR with the MusiCLEF initiative (http://ims.dei.unipd.it/websites/MusiCLEF/) [7]. The ACM Special Interest Group on Knowledge Discovery and Data Mining organizes a yearly competition, the KDD Cup, focusing on diverse Data Mining topics every year, and in 2011, the competition focused on a core MIR topic: Music Recommendation (http://www.kdd.org/kdd2011/kddcup.shtml, http://kddcup.yahoo.com/). In 2012, the MediaEval (Benchmarking Initiative for Multimedia Evaluation, see http://www.multimediaeval.org/) organizes for the first time a music-related task. Also in 2012 appears the Million Song Dataset challenge, a music recommendation challenge opened to many different sorts of data (user data, tags, etc., more details on http://www.kaggle.com/c/msdchallenge).

The establishment of an annual evaluation forum (MIREX), globally accepted by the community, and the appearance of relevant satellite forums in neighboring fields have undoubtedly been beneficial to the MIR field. However, a lot of work is still necessary to reach a level where evaluations will have a systematic and traceable positive impact on the development of MIR systems and on the creation of new knowledge in MIR. Since about 10 years, meta-evaluation methodologies have been instrumental in advancements of the Text Information Retrieval field, they need to be addressed in MIR too [8].

References

[1] D. Pearce and H. Hirsch. “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions.” In Proceedings International Conference on Spoken Language Processing, 2000.

[2] I. Guyon, S. Gunn, A. Ben Hur, and G. Dror. “Result analysis of the NIPS 2003 feature selection challenge.” In Proceedings Neural Information Processing Systems Conference, pages 545–552, 2004.

[3] M. Minsky and O. Laske. A conversation with Marvin Minsky. AI Magazine, 13(3):31–45, 1992.

[4] J. Stephen Downie (editor). The MIR/MDL Evaluation Project White Paper Collection, Edition #3 available online at http://www.music-ir.org/evaluation/wp.html, http://www.music-ir.org/evaluation/wp3/wp3_entire.pdf

[5] Cano, P. Gómez, E. Gouyon, F. Herrera, P. Koppenberger, M. Ong, B. Serra, X. Streich, S. Wack, N. ISMIR 2004 Audio Description Contest. Music Technology Group Technical Report, University Pompeu Fabra, MTG-TR-2006-02, 2006. Available online at: http://mtg.upf.edu/files/publications/MTG-TR-2006-02.pdf

[6] J. Stephen Downie. “The Music Information Retrieval Evaluation eXchange (MIREX)”, D-Lib Magazine, December 2006, volume 12 number 12

[7] Nicola Orio, David Rizo, Riccardo Miotto, Markus Schedl, Nicola Montecchio and Olivier Lartillot, MusiCLEF: a Benchmark Activity in Multimodal Music Information Retrieval. ISMIR 2011

[8] Julián Urbano. “Information Retrieval Meta-Evaluation:Challenges and Opportunities in the Music Domain”. International Society for Music Information Retrieval Conference, pp. 597-602, 2011.

3. Socio-cultural perspective

Music Information Research (MIReS) comprises research aimed at understanding and modeling music-related data in its full contextual complexity. Music is a communication phenomenon that involves people and communities immersed in specific social and cultural context. MIReS aims at processing musical data that captures the social and cultural context and at developing data processing methodologies with which to model the whole musical phenomenon.

3.1 Social aspects

3.1.1 State of the Art

Even though most of XXth Century technologies have made possible different modes of experiencing music individually, if we consider all the cultures in the world, music is still mostly experienced and valued in a social context and even in the Western culture individual listening becomes a social activity as the experience is frequently, afterwards or simultaneously, interpreted and shared with other people. Hence, the value of music as social mediator and the social dynamics it makes possible have yet to be properly addressed by researchers. In addition to a traditional view corresponding to the *social psychology of music/sociology of music* (see section 2.4) we consider two research perspectives on music social aspects: human dynamics and social computing.

How has MIR addressed, supported or capitalized on the social aspects of music? What is still to be done? As an orientation, the word "social" can be found in more or less 100 papers presented in the past 12 ISMIR editions but in most of the cases it is just a passing word, or part of a somehow shallow expression like "social tags" or "social networks". In the bunch of papers that really deal with social aspects (for example, Lee and Downie, 2004; McEnnis and Cunningham, 2007; Levy and Sandler, 2007; Fields et al., 2008; Cunningham and Nichols, 2009; Laplante, 2011), social psychology and social computing are dominant perspectives, whereas human dynamics has been, up to now, absent.

Human dynamics around music

Most of basic research on social aspects of music has focused on individuals with relation to significant groups (i.e., peers, family, gang, nation), as we have summarized above. Alternatively, social behavior can be considered globally, nearly getting rid of the individual (we cannot avoid the link to Asimov's "psychohistory"), like researchers on social animals (especially insects) usually do. A global understanding of the flow patterns of spread, influence and consumption/enjoyment of a specific musical agent or content calls for new techniques such as complex networks analysis or human dynamics (Barabasi, 2005). Our knowledge of the interplay between individual activity and social network is limited, partly due to the difficulty in collecting large-scale data that record, simultaneously, dynamical traces of individual behaviors, their contexts and social interactions. This situation is changing rapidly, however, thanks to the pervasive use of mobile phones and portable computing devices. Indeed, the records of mobile communications collected by telecommunication carriers provide extensive proxy of individual symbolic and physical behaviors and social relationships. The high penetration of mobile phones implies that such data captures a large fraction of the population of an entire country. The availability of these massive CDRs (Call Detail Record) has made possible, for instance, the empirical validation in a large-scale setting of traditional social network hypotheses (Wang et al., 2011). Taking advantage of them for music-related purposes is still pending because massive geo-temporally tagged data is still one of the bottlenecks for MIR researchers. We are still lacking of knowledge about listening patterns and how they are modulated by interaction with peers, by sharing of musical information with peers, or by geographical and environmental conditions (e.g., weather, time of the day). In order to study massive concurrent behavior patterns we only have available a big dataset of last.fm scrobblings harvested and donated by Òscar Celma. It is interesting to note that the most recent Million Song Dataset does not include any geo-temporal information. Telecommunication service companies should then be targeted by researchers and research project managers in order to make some progress along this line.

Music-related social computing

The *social computing* view, on the other hand, addresses either the creation of social conventions and context by means of technology (i.e., wikis, bookmarking, networking services, blogs), or the creation of data, information and knowledge in a collective and collaborative way (e.g., by means of collaborative filtering, reputation assignment systems, tagging (Lamere, 2008), game playing (von Ahn, 2006), collaborative music creation tools, etc.). It is usually assumed that social computation, sometimes also called social information processing, will be more effective and efficient than individual or disconnected efforts (Surowiecki, 2004). When information is created socially, it is not independent of people, but rather is significant *precisely* because it linked to people, who are in turn associated with other people (Erickson,2011). Games with a purpose (GWAP) are a paradigmatic example of social computation for annotation of different knowledge domains. Major Miner, The Listen Game, TagATune, MagnaTagATune (Law et al., 2009), Moodswings (Kim et al., 2008), Mooso, HerdIt (Barrington et al., 2009), etc., have been successfully used for gathering massive ground-truth "annotations" of music excerpts or for generating data about music preference or relatedness (see above section *Collecting music related data*). A further step in generating knowledge consists in building ontologies from tagging and writing behavior inside a delimited social network (Levy & Sandler, 2007; Pan et al., 2009). A unified model of social networks and semantics where social tagging systems can be modeled as a tripartite graph with actors, concepts and instances (e.g., songs or files) makes possible, by analyzing the relations between concepts both on the basis of co-occurrence in instances and common usage by actors (users), the emergence of lightweight ontologies from online communities (Mika,2007). A completely different approach to community knowledge extraction for the design of ontologies is the implementation of Web portals with collaborative ontology management capabilities (Zhdanova, 2008). We have recently reported on these strategies related to the Freesound community (Font et al, 2012). In addition to games and tag-related activity, musical collective knowledge can be generated by means of musical activity itself (and not just by tags or texts). Collective generation of playlists has been studied under different perspectives (Sprague et al., 2008; Stumpf & Muscroft, 2011). Precisely in this category Turntable.fm(unavailable in many European countries) is one of the recent successful musical apps for the iPhone (but see also Patent US7603352, or just the collective playlist creationg function as available in Spotify). Mashups (Sinreich, 2010) are another contemporary type of music content that benefits from music audio and context analysis technologies (Griffin et al., 2010) although it is still pending to study how collective knowledge emerges inside communities that are focused on them. To conclude, a proper multidisciplinary forum to discuss music social computation would be the "International Conference on Social computing, behavioural modeling and prediction" (held since 2008).

References

- Barabási, A.L. (2005). The origin of bursts and heavy tails in human dynamics. Nature 435, 207–211.
- Baltrunas and Amatriain, 2009

- Barrington, L, O'Malley, D., Turnbull, D. and Lanckriet, G. User-centered design of a social game to tag music. In Proceedings of the ACM SIGKDD Workshop on Human Computation, pages 7–10. ACM, 2009.
- Baur D., Butz A. Pulling Strings from a Tangle: Visual-izing a Personal Music Listening History. *Proc. of IUI 2009*, 439-444.
- Baur, D., Seiffert, F., Sedlmair, M., Boring, S, (2010). The Streams of Our Lives: Visualizing Listening Histories in Context. IEEE Transactions on Visualization and Computer Graphics, vol. 16, no. 6, pp. 1119-1128, doi:10.1109/TVCG.2010.206.
- Chai,S.K. and Salerno, J. (2010) Advances in Social Computing New York: Springer, Lecture Notes in Computer Science.
- Chen, Y., Boring, S., Butz, A. How Last.fm Illustrates the Musical World: User Behavior and Relevant User-Generated Content In Proceedings of the international workshop on Visual Interfaces to the Social and Semantic Web, VISSW 2010, Hong Kong, China.
- Cunningham, S.J. and Nichols, D. (2009): "Exploring social music behaviour: an investigation into music search at parties", Proceedings of 10th International Society for Music Information Retrieval Conference, Kobe, Japan.
- Erickson, Thomas (2011): Social Computing. In: Soegaard, Mads and Dam, Rikke Friis (eds.). "Encyclopedia of Human-Computer Interaction". Aarhus, Denmark: The Interaction-Design.org Foundation. Available online at http://www.interaction-design.org/encyclopedia/social_computing.html
- Fields, B., Jacobson, K., Rhodes, C., and Casey, M.(2008). "Social playlists and bottleneck measurements: Exploiting musician social graphs using content-based dissimilarity and pairwise maximum flow values," in Proc. of Int. Symposium on Music Information Retrieval, Philadelphia, PA, USA. pp 559-564.
- Font, F., Roma G., Herrera P., & Serra X. (2012). Characterization of the Freesound Online Community. Third International Workshop on Cognitive Information Processing.
- Griffin, G., Kim, Y. E., and Turnbull, D. (2010). Beat-syncmash-coder: A web application for real-time creation of beat-synchronous music mashups," in Proc. of the IEEE Conf. on Acoustics, Speech, and Signal Processing.
- Herrera, P., Resa Z., & Sordo M. (2010). Rocking around the clock eight days a week: an exploration of temporal patterns of music listening. 1st Workshop On Music Recommendation And Discovery (WOMRAD), ACM RecSys, 2010, Barcelona, Spain.
- Kim, Y. E., Schimdt, E., and Emelle, L. (2008). Moodswings: a collaborative game for music mood label collection. *Proceedings of the 2008 International Conference on Music Information Retrieval*, Philadelphia, PA: ISMIR.
- Lamere. P. (2008) Social Tagging and Music Information Retrieval. Journal of New Music Research: Special Issue: From Genres to Tags: Music Information Retrieval in the Age of Social Tagging, 37(2):101-114.
- Laplante, A. (2011). Social Capital and Music Discovery: An Examination of the Ties through Which Late Adolescents Discover New Music, ISMIR 2011, pp. 341-346.
- Law, E., West, K., Mandel, M., Bay, M., Downie, J. S. Evaluation of algorithms using games: the case of music tagging. *Proc. ISMIR 2009*, 387-392.

- Lee, J.H. and Downie, J.S. (2004). Survey of Music Information Needs, Uses, and Seeking Behaviours: Preliminary Findings. Proceedings of the 5th International Conference on Music Information Retrieval Barcelona, Spain. pp 441-446.
- Levy, M. and Sandler, M. (2007) A semantic space for music derived from social tags. In Proceedings of the 8th International Conference on Music Information Retrieval, Vienna, Austria.
- McEnnis, D. and Cunningham, S. J. (2007). Sociology and Music Recommendation Systems. Proceedings of 8th InternationSprague, D., Wu, F. and Tory, M. (2008). Music selection using the PartyVote democratic jukebox. In *Advanced Visual Interfaces (AVI) 2008, pp. 433-436.*
- Mika, P. (2007). Ontologies are us: A Unified Model of Social Networks and Semantics," *Journal of Web Semantics*, 5(1), pp. 5-15.
- Pan, J. Z., Taylor, S., Thomas, E. (2009) *MusicMash2: Mashing Linked Music Data via An OWL DL Web Ontology.* In: Proceedings of the WebSci'09: Society On-Line, 18-20 March 2009, Athens, Greece.
- Sinnreich, A. (2010) *Mashed Up: Music, Technology, and the Rise of Configurable Culture.* University of Massachusetts.
- Stumpf, S. and Muscroft, S. (2011). When Users Generate Music Playlists: When Words Leave Off, Music Begins?,Proc. ICME 2011, pp. 1–6.
- Surowiecki, J. (2004). The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business economies societies and nations. New York: Doubleday.
- Turnbull, D., Liu, R., Barrington, L., and Lanckriet, G. (2007). A Game-based Approach for Collecting Semantic Annotations of Music. In Proceedings of the 8th Interna- tional Conference on Music Information Retrieval (ISMIR 2007), Vienna, Austria, September.
- von Ahn, L. (2006). Games with a purpose. IEEE Computer Magazine, 39(6), 92-94.
- Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabási, A.L. (2011). Human mobility, social ties, and link prediction. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (2011), pp. 1100-1108. doi:10.1145/2020408.2020581.
- Zhdanova, A. V. Community-driven Ontology Construction in Social Networking Portals, *Web Intelligence and Agent Systems: An International Journal*, vol. 6, pp. 93-121, 2008.

3.2 Culture specificity

Most music makes very little sense unless we experience it in its proper cultural context, thus the processing of music information has to take into account this cultural context. This section deals with the state of the art and challenges of the cultural specific research issues in music information research.

3.2.1 State of the art

The current research in MIR is being done from a variety of methodologies, but the most common approximation is based on using signal processing and machine learning methods that treat musical data as any other machine readable date, thus without much domain knowledge. When we talk about Computational Musicology or Computational Ethnomusicology we are putting emphasis on the musical and cultural aspects, does incorporating domain knowledge.

The diverse cultural contexts have not been researched too much by the MIR community and we have always assumed the context of the current commercial western culture. This music context has conditioned the problems that we are working on and thus most of the solutions obtained so far. If we study other types of music and other types of cultural contexts we find new interesting problems to be solved that require new methodologies and new solutions. At the same time, working on more diverse musical repertoires, contributes to preserve the richness of our world music, which is an important mission of our research community.

When looking at the musical concepts used in MIR from a multicultural perspective most of them need to be rethought. Even clear, for most of us, concepts like tuning, rhythm, melody, … are very culture specific, and need to be treated as such. When we go into more specific terms, like scale, chord, tonic, … then it becomes even clearer. There is a need, for the musical issues prone to be studied from an MIR perspective, to be treated from a culture specific perspective. Some music traditions have fundamental differences from our most studied western music traditions, such as different musical instruments, tuning systems, performance styles, or musical forms, and that implies that at the level of feature analysis, most of the descriptors and extraction methodologies being used to analyze commercial western music are not appropriate, or at least they have to be developed much further.

[Tzanetakis et al. 2007] introduced the concept of computational ethnomusicology to refer to the use of computer tools to assist in ethnomusicological research. In their paper, they provided some ideas and specific examples of this type of multidisciplinary research and since then we have seen an increasing number of research articles related to this topic. For instance, according to [Cornelis et al. 2010], the percentage of papers on this area at the annual ISMIR conference increased from 4.8% in 2002 to 8.1% in 2008. A year later, in 2009, ISMIR hosted an oral session devoted to the analysis of folk music, sociology and ethnomusicoly. After this event, a group of researchers working on MIR and ethnomusicology started the EthnoMIR discussion group (https://groups.google.com/group/ethnocomp). Since then, the group has organized a yearly workshop on Folk Music Analysis (FMA 2011 in Athens, 2012 in Seville) with the purpose of gathering researchers who work in the area of computational music analysis of music from different cultures, using symbolic or signal processing methods, to present their work, discuss and exchange views on the topic. At the ISMIR 2011 there was also a session dedicated to "non-western music" that included two papers from a project funded by the European Research Council entitled "CompMusic: Computational Models for the discovery of the world's music." [Serra 2011]. This project is studying five art music traditions (Hindustani, Carnatic, Turkish-makam, Andalusi, and Han) from a MIR perspective.

Within the field of musicology there has been quite a bit of research applying computational methodologies. This research is normally referred as Computational Musicology [Camilleri 1993] and within it there is a strong emphasis on the processing of the musical notation of the western classical music. A good source of references is the publication Computing in Musicology [Hewlett & Selfridge-Field].

For the specific case of non-western music there has been some work based on the study of different facets such as timbre/instrumentation (e.g. [Proutskova & Casey 2009]), rhythm (e.g. [Holzapfel & Stylianou 2009]), motives (e.g. [Lartillot & Ayari 2006] [Conklin & Anagnostopoulou 2011]), tuning and scale (e.g. [Gedik & Bozkurt 2009] [Moelants et al. 2009]), melody (e.g. [Wiering & al. 2009] [Mora et al. 2012]) or performance variations (e.g. [Müller et al. 2012]).

References

- [Camilleri 1993] Camilleri, L. "Computational Musicology A Survey on Methodologies and Applications", Revue Informatique et Statistique dans les Sciences humaines, 1993.
- [Conklin & Anagnostopoulou 2011] Conklin, D., Anagnostopoulou, C. “Comparative pattern analysis of Cretan folk songs”. *Journal of New Music Research*, 40(2):119-125, 2011.
- [Cornelis et al. 2010] Cornelis, O., Lesaffre, M., Moelants, D., Leman, M. "Access to ethnic music: advances and perspectives in content-based music information retrieval". Signal Processing, 90, pp. 1008-1031. 2010.
- [Gedik & Bozkurt 2009] Gedik, A. C., Bozkurt, B., "Evaluation of the Makam Scale Theory of Arel for Music Information Retrieval on Traditional Turkish Art Music", Journal of New Music Research, 38:2,pp. 103-116, 2009.
- [Hewlett & Selfridge-Field] Hewlett, W.B., Selfridge-Field, E. eds. Computing in Musicology (Menlo Park CA: CCARH).
- [Holzapfel & Stylianou 2009] Holzapfel, A., Stylianou, Y,. "Rhythmic Similarity in Traditional Turkish Music", *ISMIR 2009*.
- [Lartillot & Ayari 2006] Lartillot, O., Ayari, M., "Motivic pattern extraction in music, and application to the study of Tunisian Modal Music", *South African Computer Journal*, 36, pp. 16-28, June 2006.
- [Moelants et al. 2009] Moelants D., Cornelis O., Leman M., Exploring African tone scales, Proceedings ISMIR (2009), Kobe, Japan.
- [Mora et al. 2012] Mora, J., Gomez, F., Gómez, E., Escobar-Borrego, F.J., Diaz-Bañez, J.M.Melodic Characterization and Similarity in A Cappella Flamenco Cantes. 11th International Society for Music Information Retrieval Conference (ISMIR 2010) .
- [Müller et al. 2012] Müller, M., Grosche, P. & Wiering, F. (2010). Automated Analysis of Performance Variations in Folk Song Recordings. In *Proceedings 11th ACM SIGMM International Conference on Multimedia Information Retrieval (ACM MIR).*
- [Tzanetakis et al. 2007] Tzanetakis, G. et al.: "Computational Ethnomusicology," Journal of Interdisciplinary Music Studies, 1(2), pp. 1-24, 2007.
- [Proutskova & Casey 2009] Proutskova, P., Casey, M. You Call That Singing? Ensemble Classification for Multi-Cultural Collections of Music Recordings, ISMIR 2009.
- [Serra 2011] Serra, X. "A Multicultural Approach in Music Information Research." *Proceedings of the Int. Soc. for Music Information Retrieval Conf.* 2011.
- [Wiering & al. 2009] Wiering, F., Veltkamp, R.C., Garbers, J., Volk, A., Kranenburg, P. van & Grijp, L.P. (2009). Modelling Folksong Melodies. *Interdisciplinary Science Reviews, 34*(2-3), 154-171.

3.3 User behaviour

One of the relevant views on music research is to take the user (performer or listener) perspective, given that the user is central to any music experience. Here we overview the research and challenges to MIR research from the user perspective.

3.3.1 State of the art

Activities related to Western commercial music can be grouped into

- listening (to recorded media or live performances; review/discussion of what was heard)
- performing (interpretation, improvisation, rehearsal, recording, live performance) and
- creating (composition, recording, studio production, improvisation),

Within each group, MI research can relate to the analysis of practices or to the proposal of tools to help the practice.

A. Listening

Among these categories, research presented in conferences such as ISMIR mainly focus on the listening scenario: propose tools to help people access (listen to) music. But little attention is paid to analyzing user practices. As pointed out by [Weigl and Guastavino, 2011], a focus on the user has repeatedly been identified as a key requirement for future MIR research, yet empirical user studies have been relatively sparse in the literature, the overwhelming research attention in MIR remaining systems-focused. Important questions are: What are the user requirements and information needs? How do people organize their music? How would they like to see, access, search over digital libraries? What is the influence of the listening context? What is the role of social relations? Actually given that (one of) the Grand-Challenges in MIR is the creation a full-featured system [Downie et al., 2009], these questions should be answered in order to make the system useful for users. This is especially true considering that the results provided by the few research in the subject provided unexpected results. For example [Laplante and Downie, 2006] showed that part of the users are seeking new music without specific goals in mind, just for updating and expanding their musical knowledge and for the pleasure of searching. With this in mind, systems should therefore support various browsing approaches. [Cunningham et al., 2004] highlight user needs for use tagging (scenarios in which a given piece of music might be relevant), a subject currently largely under-studied. [Laplante, 2010] identifies the changes in musical taste according to social factors and [Cunningham and Nichols, 2009] suggest support for collaborative play-list creation. [Uitdenbogerd and Yap, 2003] conclude that textual queries for melodic content are too difficult to be used by ordinary users. According to [Kolhoff et al., 2008], landscape representations or geographic views of music collections have certain disadvantages and that users seem to have preferences for simple and clean interfaces. A recent survey made within the Chorus+ EU project [Lidy and van der Linden, 2011], also highlights important point such as the prevalence of YouTube as the most-used music service (among participants to the survey). Also it highlights the fact that most people search using artist, composer, song title, album or genre but the search possibilities enabled by new technologies (taste, mood or similarity) appear less prevalent.

B. Performing

If few papers relate to the listener-behaviour, this is not the case for performers and performances (in terms of music concerts, opera, theatre, dance) or interactions (interactive installations or instruments). A large community has been studying the subject of performance from the pioneer works of [Seashore, 1938]. In this, a performer is considered as the essential mediator between composer and listener. These studies show the impact of the performer, the performances, the large-structure and micro-structure, and the intentional mood on the choice of tempo, timing, loudness, timbre and articulation [Rink, 1995], [Gabrielsson, 2003]. First experiments were made using piano analysis (for ease of event-recoding) [Parncutt, 2003], but today they are extended to saxophone [Ramirez et al., 2007], cello [Chudy and Dixon, 2010] and singing voice. Understanding the process of performance has several goals: a better understanding of what makes a great interpretation (the Horowitz or Rachmaninov factors [Widmer et al., 2003]); music education; and automatic expressive perfomances (KTH model of [Sundberg et al., 1983] and Rendering Contest (Rencon)). Tools to visualize performance interpretation have also been proposed [Dixon et al., 2002].

According to [Delgado et al., 2011], different research strategies can be distinguished: (a) analysis-by-measurement (based on acoustic and statistical analysis of performances); (b) analysis-by-synthesis (based on interviewing expert musicians); and (c) inductive machine learning applied to large database of performances. The latter is the most closely related to current MIR research, which can be of great help for this (see [Chudy and Dixon, 2010]). Considering that performance is not limited to the instrumentalists, the conductor is also studied [Luck et al., 2010], and research includes studies on interaction and gesture ([Jorda, 2003], [Bevilacqua et al., 2011]). The large number of related contributions at conferences such as ISPS (International Symposium on Performance Science) shows that this domain is very active. As another example of the activity in this field, the current SIEMPRE EU project aims at developing new theoretical frameworks, computational methods and algorithms for the analysis of creative social behaviour with a focus on ensemble musical performance.

C. Composing

While historical musicology aims at studying composition once published, hence not considering the composition practice, new projects such as MuTec2 aim at following composers during their creative project (using sketches, drafts, composer interviews, and considering composer readings). Related to this new field, the conference TCPM-2011 ”Tracking the Creative Process in Music” has been created. The group of Barry Eaglestone [Nuhn et al., 2002] at the Information Systems and the Music Informatics research groups also studies the composer practices.

References

- [Bevilacqua et al., 2011] Bevilacqua, F., Schnell, N., and Alaoui, S. (2011). Gesture capture: Paradigms in interactive music/dance systems. In Verlag, T., editor, Emerging Bodies: The Performance of Worldmaking in Dance and Choreography, page 183.
- [Chudy and Dixon, 2010] Chudy, M. and Dixon, S. (2010). Towards music performer recognition using timbre. In Third International Conference of Students of Systematic Musicology.
- [Cunningham et al., 2004] Cunningham, S., Jones, M., and Jones, S. (2004). Organizing digital music for use: an examination of personal music collections. In Proc. of ISMIR, pages 447– 454, Barcelona, Spain.
- [Cunningham and Nichols, 2009] Cunningham, S. and Nichols, D. (2009). Exploring social music behaviour: An investigation of music selection at parties. In Proc. of ISMIR, pages 26–30, Kobe, Japan.

- [Delgado et al., 2011] Delgado, M., Fajardo, W., and Molina-Solana, M. (2011). A state of the art on computational music performance. Expert Systems with Applications, 38(1):155–160.
- [Dixon et al., 2002] Dixon, S., Goebl, W., and Widmer, G. (2002). The Performance Worm: Real time visualisation of expression based on Langner's tempo-loudness animation. In Proc. of ICMC.
- [Downie et al., 2009] J.S. Downie, D. Byrd and T. Crawford (2009). Ten Years of ISMIR: Reflections on Challenges and Opportunities. 10th International Society for Music Information Retrieval Conference, 13-18.
- [Gabrielsson, 2003] Gabrielsson, A. (2003). Music performance research at the millennium. Psychology of music, 31(3):221–272.
- [Jorda, 2003] Jorda, S. (2003). Interactive music systems for everyone exploring visual feedback as a way for creating more intuitive, efficient and learnable instruments. In SMAC.
- [Kolhoff et al., 2008] Kolhoff, P., Preuß, J., and Loviscach, J. (2008). Content-based icons for music files. Computers & Graphics, 32(5):550–560.
- [Laplante, 2010] Laplante, A. (2010). The role people play in adolescents music information acquisition. In Proc. of WOMRAD (Workshop on Music Recommendation and Discovery).
- [Laplante and Downie, 2006] Laplante, A. and Downie, J. (2006). Everyday life music information-seeking behaviour of young adults. In Proc. of ISMIR, pages 381–382, Victoria, Canada.
- [Lidy and van der Linden, 2011] Lidy, T. and van der Linden, P. (2011). Think-tank on the future of music search, access and consumption. In (FP7/2007-2013), E. C. S. F. P., editor, MIDEM, Cannes, France.
- [Luck et al., 2010] Luck, G., Toiviainen, P., and Thompson, M. (2010). Perception of expression in conductors' gestures: A continuous response study. Music Perception, 28(1):47–57.
- [Nuhn et al., 2002] Nuhn, R., Eaglestone, B., Ford, N., Moore, A., and Brown, G. (2002). A qualitative analysis of composers at work. In Proceedings of the International Computer Music Conference, Gothenburg, International Computer Music Association, pp597-599. Citeseer.
- [Parncutt, 2003] Parncutt, R. (2003). Accents and expression in piano performance. Perspektiven und Methoden einer Systemischen Musikwissenschaft, pages 163–185.
- [Ramirez et al., 2007] Ramirez, R. et al. (2007). Identifying saxophonists from their playing styles. Proceedings of the 30th AES International.
- [Rink, 1995] Rink, J. (1995). The Practice Of Performance: Studies In Musical Interpretation. Cambridge University Press Cambridge.
- [Seashore, 1938] Seashore, C. E. (1938). Psychology of music. New York: McGraw-Hill.
- [Sundberg et al., 1983] Sundberg, J., Fryden, L., and Askenfelt, A. (1983). What tells you the player is musical? An analysis-by-synthesis study of music performance, volume 39, pages 61–75. Stockholm, Sweden: Publication issued by the Royal Swedish Academy of Music.
- [Uitdenbogerd and Yap, 2003] Uitdenbogerd, A. and Yap, Y. (2003). Was parsons right? an experiment in usability of music representations for melody-based music retrieval. In Proc. of ISMIR, pages 75–79, Baltimore, Maryland, USA.
- [Weigl and Guastavino, 2011] Weigl, D. and Guastavino, C. (2011). User studies in the music information retrieval literature. In Proc. of ISMIR, Miami, USA.
- [Widmer et al., 2003] G. Widmer, S. Dixon, W. Goebl, E. Pampalk and A. Tobudic (2003). In Search of the Horowitz Factor. AI Magazine, 24 (3), 111-130.

4. Exploitation perspective

Music Information Research (MIReS) comprises research aimed at producing exploitable technologies for storing, organising, retrieving, delivering, displaying and connecting information related to music. These technologies will enable novel user experiences, commercially successful large-scale applications, services and distribution channels for players in the digital media industry.

4.1 Music Industry Applications

4.1.1 State of the art

In reviewing the state of the art in music industry applications we do not aim for a complete enumeration but instead concentrate on a number of issues and aspects we identified as being vital for the future of electronic music distribution. This includes “Search and discovery” (4.1.1.1) in music databases as one of the applications with a large industrial potential that is already quite developed. We also focus on the often neglected but very important “Interface aspects” for industry and commercial applications (4.1.1.2). Finally we review (4.1.1.3) how “Music Rights” issues influence music industry applications.

4.1.1.1 Search and Discovery

As could be witnessed over the last years, music is being produced and published at a faster rate than any individual could actually listen to it: estimates range form yearly 11,000 (nonclassical) major label albums averaging some ten songs per album ([Vogel 2004], p. 261) up to 97,751 albums released in the United States in 2009, as reported by Nielsen SoundScan (http://www.billboard.biz/bbbiz/content_display/industry/news/e3i4ad94ea6265fac02d4c813c0b6a93ca2). The corpus of available music is growing at an extraordinary rate of possibly over five hours of newly released music per hour. As a consequence, any music listener has to rely on preselections to provide an appropriate choice of music she or he likes. A person searching for new music traditionally only had limited influence on these preselections, such as choosing a radio station to listen to or a shelve in a music store to look through, and was largely dependend on other people's choices, such as the station's editorial decisions on the playlist, or the selection and sorting of records by staff in the store.

Digital technologies have changed this situation in at least two respects: digital music distribution channels such as iTunes, Amazon or Spotify can provide quick access to millions of music pieces at very low cost, hence they are less strictly preselected, and, with the abandonment of physical records, they shifted granularity from albums to single tracks, making it even harder for potential customers to make a choice. To fill this gap of missing preselections, automatic music recommendation systems supporting search and discovery have been developed attempting to provide an improved and manageable access to the music of the world.

In what follows, we will introduce and discuss the state of the art in music recommendation by reviewing a number of popular music-related Internet platforms which have received wide attention in the general public. Many of these industry applications have their roots in the MIR community and serve as successful examples of how exploitation of our results may work in the future.

Amazon (www.amazon.com) suggests albums or songs based on what has been purchased in the same order or by the same customers as items one searched for or bought. This is a form of collaborative filtering [Herlocker et al 1999], which assumes that users who have agreed in the past (in their purchase decisions) will also agree in the future (by purchasing the same items). Collaborative filtering generally suffers from two related problems: the coldstart problem and the popularity bias. The coldstart problem is the fact that albums that have not yet been purchased by anybody can never be suggested. The popularity bias is the problem that for any given item, popular albums are more likely to have been purchased in conjunction with it than unpopular ones, and so have a better chance of being recommended. In consequence, collaborative filtering alone is incapable of suggesting new music releases. An additional problem specific to Amazon is that users may purchase items for somebody else (e.g., as a present), which might flaw the recommendations generated both for them and for other users of allegedly the same taste. Spotify (www.spotify.com), a music streaming service, bases its recommendations on its users' listening behavior, analyzing which artists are often played by the same listeners. While this may potentially result in better suggestions than analyzing sparse data such as record purchases, it is again subject to the cold-start problem and popularity bias. Furthermore, Spotify only recommends related artists and not songs, which is rather unspecific. Genius is a function in Apple iTunes (www.apple.com/itunes/) which generates playlists and song recommendations by comparing music libraries, purchase histories and playlists of all its users, possibly integrating external sources of information. Assuming such external information does not play a major role, this system is again based mainly on collaborative filtering. Last.fm combines information obtained from users' listening behavior and user-supplied tags (words or short expressions describing a song or artist). Tags can help making recommendations transparent to users, e.g. a user listening to a love song may be recommended other tracks that have frequently been tagged as 'slow' and 'romantic'. But they are also inherently erroneous due to the lack of carefulness of some users, which requires a range of counter measures for data cleaning. And of course tags are also affected by the cold-start problem and popularity bias. Pandora (www.pandora.com), another music streaming service, recommends songs from its catalog based on expert reviews of tracks with respect to a few hundred genre-specific criteria. This allows very accurate suggestions of songs that sound similar to what a user listens to, including sophisticated explanations for why a song was suggested (e.g., a track may be recommended because it is in 'major key', features 'acoustic rhythm guitars', 'a subtle use of vocal harmony' and exhibits 'punk influences'). Such expert reviews incur high costs in terms of time and money which makes it impossible to extend the catalog at a rate that can keep up with new releases. This has a limiting effect on the selection of music available to users.

Two more general music service providers with its roots firmly in the field of MIR have appeared in the recent years: “The Echo Nest” (http://the.echonest.com/) and “BMAT” (http://www.bmat.com). They both provide music services to both developers and media companies via an API (application programming interface). This is a whole new and very important business model which is already servicing hundreds of applications already. Its services range from recommendation, playlisting, fingerprinting to general audio analysis. As an example, BMAT services Samsung’s MusicHub or 247’s Juke amongst others.

Most approaches described so far rely on some form of meta-information: user's listening or purchasing behavior, statistics about artists and genres in music collections, user defined tags etc. Another option is to actually analyse the audio content trying to model what is important for the perceived similarity between songs: instrumentation, tempo, rhythm, melody, harmony, etc. While many research prototypes of recommendation systems that use content-based audio similarity have been described in the literature (e.g., [Pampalk 2001], [Neumayer et al 2005], [Lamere & Eck 2007], [Knees et al 2007], to name just a few), very little has been reported about successful adoption of such approaches to real-life scenarios. A music recommender that supports exploration of a national data base for amateur and up-and-coming artists has been reported (www.soundpark.at, [Gasser & Flexer 2009]). Mufin (www.mufin.com) is advertised as a music discovery engine that uses purely content-based methods. MusicIP (www.musicip.com) offers the 'Mixer'-application that uses a combination of content-based methods and metadata to generate playlists. BMAT's Ella (http://www.bmat.com/products/ella/index.php) offers a hybrid music recommender with leverages music content analysis, tag similarity and collaborative filtering methods.

An application of MIR research that is somewhat related to search and discovery is that of audio fingerprinting. Here the aim is to identify one specific song based on an often noisy and incomplete audio recording. While, for the basic use cases, the research and engineering problems of audio fingerprinting have practically been solved; for other real industry use cases, the research in this area is still trying to solve background music detection (over voice), on noisy backgrounds and with edited music. For end users, the Shazam-service (mainly for mobile phones, http://www.shazam.com) is very successful and dominating the market. In the business to business segment, a number of players share the market: BMAT in Spain, Tunesat and mediaguide in the USA, mediaforest in Israel, Nielsen and kollector in Europe, Monitec in Southamerica, Soundmouse in the UK and yacast in France.

A very early example of audio hardware containing MIR technology was the series of Thomson RCA Lyra media players (http://en.wikipedia.org/wiki/RCA_Lyra). At that time, one of the main obstacles to successful commercialization was the necessity to analyse the music files offline due the lack of computational power of portable hardware. A couple of rare examples of audio equipment containing MIR technology that can be bought in stores today is the Beosound 5 (www.beosound5.com) home entertainment center released in 2009 by Bang&Olufsen and Yamaha's Bodibeat (www.yamaha.com/bodibeat) released in 2007. The former integrates content-based audio similarity with a simple 'More Of The Same Music'- user interface, that allows users to create playlists by choosing an arbitrary seed song. The latter includes playlist generation based on descriptors such as tempo.

A major challenge a new technology must face when it is to be applied in viable commercial products is scalability; that is the ability of the technology to handle massive amounts of data and the ability to handle that data's eventual growth in a cost effective manner. The problem is twofold. Firstly, some techniques are simply neither deployed nor tested since it's computationally impossible due to the size of datasets. Secondly, assuming the technique is scalable from a non-functional point of view, applying it to multi-million datasets may reveal problems which were not obvious in the first place. Beyond the problem of handling 'big data', granting the research access to huge music related datasets may generate beneficial by-products to the music information research world. First, in large collections, certain phenomena may become discernible and lead to novel discoveries. Secondly, a large dataset can be relatively comprehensive, encompassing various more specialized subsets. By having all subsets within a single universe, we can have standardized data fields, features, etc. Lastly, a big dataset available to academia greatly promotes the interchange of ideas and results leading to, yet again, novel discoveries. A good example here is the "Million Songs Dataset" (http://labrosa.ee.columbia.edu/millionsong) which contains user tags provided by Last.Fm.

Systems that are able to automatically recommend music (as described above) are one of the most commercially relevant outcome from the MIR community. For such recommender systems it is especially important to being able to work with very large collections of music. The core technique driving automatic music recommendation systems is the modeling of music similarity which is one of the central notions of MIR. Proper modeling of music similarity is at the heart of every application allowing automatic organization and processing of music data bases. Scaling up computation of music similarity to the millions is therefore an essential concern of MIR. Scalable music recommendation systems have been the subject of a number of publications. Probably one of the first content-based music recommendation system working on large collections (over 200.000 songs) was published by [Cano et al 2005]. Latest results (see e.g. [Schnitzer et al 2012]) enable systems to answer music similarity queries in about half a second on a standard desktop CPU on a collection of 2.5 million music tracks.

The issue of scalability clearly also affects other areas of MIR: Music Identification meaning both pure fingerprinting technologies and cover detection, Multimodal Music Recommendation and Personalization (using contextual and collaborative filtering Information). Moreover, it not only affects the Retrieval stage but also the feature extraction side of the equation.

4.1.1.2 Interface Aspects

While the interactive aspect of most commercial library music applications has resorted to the metaphor of spreadsheets (e.g. *iTunes*) or rely on searching for music by filling a set of forms and radio buttons (e.g. *Synchtank*), MIR offers new opportunities for music interfaces that rely on the music itself to aid the listener in organising and finding items in digital collections. Interactive audio-visual systems utilising music information with fast feedback audio loops, where concentration is not devoted to how to operate the system but instead focused on the quality and features of the content, can contribute to more efficient communication.

Several examples from the research conducted by the Music Information Retrieval community over the past 10 years can be seen as precursors to the current state of the art in the development of interfaces for commercial music search and discovery: early innovative approaches to visually mapping sound clusters in *Islands of Music* [Pampalk, 2001]; the launch of interfaces for collaborative projects such as *The Freesound Project* in 2005 (http://www.freesound.org/); the influence of tangible tabletop interfaces for new musical experiences like the *Reactable* [Jordà et al, 2005]; the deployment of tangible music information interfaces [Julià and Jordà 2009]; spatial audio search environments for social networking in real time in *decibel 151* [Magas et al, 2009]; and public art installations which use physical objects to interact with music information (e.g. the *Barcelona Magic Fountain* at Montjuic). New challenges are presented by the changing landscape of networked media systems, including location-based devices, web-mediated social networks, dynamic context-driven user communities and open environments.

While targeted search is a priority for commercial applications, methods deployed in research-based visual interfaces for MIR have predominately focused on visualising numerical aspects of the music, as analysed algorithmically, reflecting the mode of the scientific enquiry used to extract the data, rather than its more qualitative aspects that might be perceived by a music expert or a listener. The user controls often referenced control panels used in engineering (*MusicBox* [Lillie, 2008]) or gaming platforms

(*Musicream* [Goto and Goto, 2005]). By emphasising visualisation of research data over usability, such systems often proved a challenge for commercial music providers focused on fast and efficient targeted searches and frequent downloads. In some systems classification has been based on semantic categorization into genres or moods, aided by colour-coding to make them more accessible to the novice or non-technical user/researcher (e.g. *Music Plasma* or *Live Plasma* [Vavrille, 2005], later developed into *Musicovery* in conjunction with Pandora), making them more appropriate for commercial deployment, though often resulting in a user perception of a subjective value system.

In existing commercial systems semantic clues most frequently utilise genre or mood classification, reduced and simplified for reasons of collection management. Crowd-sourced tagging has become increasingly popular, though the most statistically popular tags tend to be generic and therefore result in simplified applications. The visual/interactive aspect of commercial library music applications has resorted to the metaphor of spreadsheets, forms and radio buttons. In such systems delays in communication often cause breakdowns in the work flow. Enhanced user experiences are offered by music information interfaces which rely on the music itself to aid the listener in organising and finding items in digital collections: more efficient communication is achieved when interactive audio-visual systems utilise music information via fast feedback audio loops, focused on the quality and features of the content (e.g. *Sonaris* and its precursor *mHashup* [Magas et al 2008]).

4.1.1.3 Music Rights

In a landscape where the music industry is facing difficult times with income from physical sales shrinking, the music rights revenues are increasing worldwide. According to [CISAC 2012] the author's society royalty collections were €7.5 billion in 2010 (climbing a 5,5% year-on-year) and [IFPI 2012] announced that the global performance rights reached the 905 US$ millions in 2011 (an increase of 4,9% from the previous year). These positive numbers are due to the increase of the number of media paying royalties and an improvement of the collecting reach of these societies.

As its name explains music rights means paying the owners of these rights (authors, performers, labels…) for the usage of the music they have created and performed [The American Society] . Those who pay more music royalties are television, radio stations and those industries whose services are based on music, like clubs or venues. Apart from those, they also pay for the music rights a lot of other companies and associations from shops or dentists to school plays, basically anyone who aims at using somebody else's music creation [Broadcast Music 2012] . In recent years the music rights revenues coming from the digital world have also grown in importance[The New York Times 2012] . All these rights are collected through the royalty collection societies, which are divided in three kinds depending on the rights they represent: Authors, Performance or Master. Most of authors' societies worldwide are associated in CISAC (www.cisac.org) while the master societies are associated in IFPI (www.ifpi.org). The societies collect music rights and distribute them among their associates. At this point a lot of controversy arises due to the different processes they use for such distribution and questions are raised about how to make it as fair as possible [Younison 2012].

Ideally every right owner should be paid for the use of their music but in practice it is difficult and expensive to control all the media and all potential venues where music could eventually be used. The solutions that have been found vary depending on the country, the society and the type of source. Some years ago, the societies used to distribute based on the results of the top selling charts which created huge inequalities between artists. Later some other systems and technologies appeared:

-Cue sheets: Media companies are obliged to fill cue sheets, the list of music broadcasted, explaining their use. However, this tends to be inaccurate because, while generating the cue sheets represents lots of work, media companies don't benefit from the accuracy of those. [Sealove]

-Watermarking: It consists in embedding an extra signal into a digital music work so this signal can be detected when the work is reproduced. Watermarking requires the use of watermarked audio references when broadcasting which is very rare. Also, the extra signal can easily be removed from original audio. [Music Trace Watermarking]

-Fingerprinting: It consists in an algorithm that extracts the main features of an audio piece making a so-called fingerprint of the track. This fingerprint may easily be matched against an audio database which may comprise recordings from television, radio or internet radio broadcasts. [Music Trace Fingerprinting]

-Clubs: The collecting societies track music played in all types of venues by sending a specialist who recognizes music and writes down a cue sheet or by installing recording stations in Dj boards. [Bemuso]

-Online: Some of the most used music channels in internet as streaming or peer to peer services are extremely difficult to monitor. Nowadays the music monitor online is based on crawling millions of webs and detect their music usage. [Skates 2012]

As we can observe a lot of technologies have been developed to increase transparency and fairness in recent years but there is still a lot of field for improvement in the music rights business.

References

- [Bemuso] Bemuso. "Music royalty collection societies". http://www.bemuso.com/musicbiz/musicroyaltycollectionsocieties.html
- [Broadcast Music 2012] Broadcast Music, Inc. "Royalty Policy Manual." Last updated 04.18.2012. http://www.bmi.com/creators/royalty/how_we_pay_royalties/detail
- [Cano et al 2005] P. Cano, M. Koppenberger, and N.Wack. An industrial-strength content-based music recommendation system. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'05, pages 673-673, New York, NY, USA, 2005.
- [CISAC 2012] CISAC. "Royalty Collections Climb to New Peak at €7.5 billion." 30th January 2012. http//www.cisac.org/CisacPortal/initConsultDoc.do?idDoc=22994#pressrelease
- [Gasser & Flexer 2009] M. Gasser, A. Flexer. FM4 Soundpark: Audio-based Music Recommendation in Everyday Use, in Proceedings of the 6th Sound and Music Computing Conference (SMC'09), Porto, Portugal, 2009.
- [Goto and Goto 2005] M. Goto, T. Goto: Musicream: New Music Playback Interface for Streaming, Sticking, Sorting, and Recalling Musical Pieces. ISMIR 2005: 404-411
- [Herlocker et al 1999] J.L. Herlocker, J.A. Konstan, A. Borchers, and J. Riedl. An Algorithmic Framework for Performing Collaborative Filtering. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), pages 230-237. ACM, 1999.

- [IFPI 2012] IFPI. “Recording Industry in Numbers – 2012 Edition”.
- [Jorda et al 2005c] S. Jordà, M. Kaltenbrunner, G. Geiger, R. Bencina, The reacTable. Proceedings of the International Computer Music Conference (ICM05).
- [Julià and Jordà 2009] Julià, C. F., and Jordà, S. (2009). Songexplorer: A tabletop application for exploring large collections of songs. ISMIR 2009.
- [Knees et al 2007] P. Knees, M. Schedl, T. Pohle, G. Widmer. Exploring Music Collections in Virtual Landscapes. IEEE MultiMedia, volume 14, number 3, pp. 46-54, July-September 2007.
- [Lamere & Eck 2007] P. Lamere, D. Eck. Using 3d visualizations to explore and discover music, Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07), 2007.
- [Lillie 2008] A. S. Lillie. Musicbox: Navigating the space of your music. Master’s thesis, Massachussets Institute of Technology, 2008.
- [Magas et al 2008] M. Magas, M. Casey, and C. Rhodes. mHashup: fast visual music discovery via locality sensitive hashing. In ACM SIGGRAPH 2008 New Tech Demos .
- [Magas et al 2009] M. Magas, R. Stewart, B. Fields, B.. decibel 151: Collaborative Spatial Audio Interactive Environment. In ACM SIGGRAPH. 2009.
- [Music Trace Fingerprinting] Music Trace. “Audio Fingerprinting”.http://www.musictrace.de/technologies/fingerprinting.en.htm
- [Music Trace Watermarking] Music Trace. “Audio Watermarking”.http://www.musictrace.de/technologies/watermarking.en.htm
- [Neumayer et al 2005] R. Neumayer, M. Dittenbach, A. Rauber. Playsom and pocketsomplayer: Alternative interfaces to large music collections. In Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR'05), 2005.
- [Pampalk 2001] E. Pampalk. Islands of music: Analysis, organization, and visualization of music archives. MSc thesis, Vienna University of Technology, 2001.
- [Schnitzer et al 2012] D. Schnitzer D., A. Flexer, G. Widmer. A Fast Audio Similarity Retrieval Method for Millions of Music Tracks, Multimedia Tools and Applications, Volume 58, Number 1, 23-40, 2012.
- [Sealove] C. Sealove. “Music Cue Sheets & Performance Royalties”. http://www.editorsguild.com/v2/magazine/Newsletter/SepOct03/sepoct03_music_cue_sheets.html
- [Skates 2012] S. Skates. Music Row. “TuneSat Debuts Exclusive Internet Monitoring Technology”. http://www.musicrow.com/2012/01/tunesat-debuts-exclusive-internet-monitoring-technology/
- [The American Society] The American Society of Composers, Authors and Publishers. “ASCAP Licensing FAQ”. http://ascap.com/licensing/licensingfaq.aspx
- [The New York Times 2012] The New York Times. “Royalties for Streaming Music Grew 17% in 2011.” 01.17.2012. http://mediadecoder.blogs.nytimes.com/2012/01/17/digital-notes-royalties-for-streaming-music-grew-17-in-2011/
- [Vavrille, 2005] F. Vavrille. Live Plasma. http://www.visualcomplexity.com/vc/project.cfm?id=123
- [Vogel 2004] H. L. Vogel. Entertainment Industry Economics: A Guide for Financial Analysis. Cambridge University Press, 6 edition, 2004.
- [Younison 2012] Younison. Artists, stand up for your rights! “Transparency & accountability in collective rights management.” 08.01.2012. http://www.younison.eu/downloads/get/23

4.2 Artistic applications

Although MIR did arguably not start as a research discipline for promoting creativity or music performance, this trend has begun to gain importance in recent years. The possibilities of MIR for supporting musical creation (e.g *novel musical instruments for music performance*) or for inspiring the creation of art pieces (e.g. i*nteractive installations*), are indeed many-folded. Apart from this first "musical instrument vs. interactive installation" dichotomy, which will be further developed in the following two subsections, the creative possibilities of MIR could be studied and classified according to many different criteria, such as *off-line tools* (for composition or editing) vs. *real-time tools* (for performance), or tools designed for *expert performers* vs. tools designed for *absolute novice* users, a later category which could include all type of applications promoting different models of "active listening". Following, we describe some of the possibilities and the achievements gained along these different lines; this list does not aim to be exhaustive or complete but rather to give an overview of the different directions that are currently being explored.

4.2.1 State of the art

Music Performance applications

Some commercial products exist that employ interactive MIR techniques. These include Celemony's Melodyne[5] and Roland's R-Mix[6], which provide studio production tools for pitch recognition and correction, tempo and timing alteration and spectrum visualisation. Although all these products allow working interactively, they are designed for off-line work (i.e. composing, editing) and not for music performance. As a matter of act, while most of the currently available MIR applications or libraries are still not meant for a real-time use, in the last years MIR has started to show its potential in this domain.

Currently, one of the more obvious MIR applications for real-time music creation is that of "concatenative synthesis" [Schwarz 2005; Maestre et al. 2009], "musaicing" [Zils and Pachet 2001] or mashup. These three terms approximately relate to the same idea, creating new music by means of concatenating short fragments of sound or music recordings to "approximate" the sound of a target piece. More precisely, an existing music piece or musical fragment is substituted with small, similar sounding music fragments, leading to a similarly structured result. The duration of these sound units can vary depending on the techniques employed and the desired aesthetic results, but are roughly in the range of 10 milliseconds up to several seconds or several musical bars. While manual procedures could used with longer fragments (i.e. of several seconds), the use of shorter fragments inevitably leads to automatised MIR analysis and recovery techniques, in which a "target" track or sound is analysed, its descriptors extracted for every small fragment, and these fragments substituted with the best candidates from a large database of sound snippets. When using a pre-analysed sound repository and a compact feature representation, these techniques can be efficiently applied in real-time. [Janer & de Boer 2008] describe a method for real-time voice-driven audio mosaicing synthesis. BeatJockey [Molina et al. 2011] is a system aimed for DJs, which integrates audio mosaicing, beat-tracking and machine learning techniques and brings them into the Reactable musical tabletop. Several commercial tools following this approach (such

[5] http://www.celemony.com/cms/index.php?id=products_editor

[6] http://www.rolandconnect.com/product_2011-09.php?p=r-mix

as Steinberg's Loopmash[7] VST plugin and iOS app) are also already available. These techniques bring new creative possibilities somewhere in between synthesis control and remixing, and open the path to radically novel control interfaces and interaction modalities for music performance.

At a slightly larger time scale, MIR techniques can also be very convenient, for instance for finding suitable loops or sound files to fit a particular composition or mix. These techniques are also more necessary as the availability of large public and free sound databases and repositories such as Freesound[8], are becoming mainstream. Using such repositories and APIs such as EchoNest's Remix Python API or MTG's Essentia[9], developers and hackers are creating a panoply of imaginative remix applications, many of them being developed during Music Hack Day events, which lately appear to be a very productive place for MIR based creation. Combining audio retrieval, audio processing and audio playback with gestural control, [Schnell et al. 2011] propose a system for the gestural re-embodiment of recorded sound and music, demonstrating a large variety of different ”playing techniques” in musical performance, using wireless motion sensor modules in conjunction with gesture analysis and real-time audio processing components. In the "Urban Music Game" installation, they implement a ball-controlled sonic game in which the different movements resulting from the passing of a ball, modifies the sounds previously associated with the ball.

Whereas *musaicing* or remixing applications, do mostly rely on low-level signal processing analysis, the following examples depict a more musical knowledge and understanding. In that sense, their MIR component has to do more with the "new" R of "Research" than with the former one of "Retrieval". [Assayag et al. 2006] describes a multi-agent architecture for an improvisation oriented musician-machine interaction system that learns in real-time from human performers, and establishes improvisatory dialogues with the performers by recycling their own audio material. The Wekinator [Fiebrink 2011] is a real-time machine learning toolkit that can be used in the processes of music composition and performance, as well as to build new musical interfaces. Pachet is working with Constrained Markov Models (CMM) for studying musical style by analyzing musicians, extracting relevant features and modeling them using CMM [Pachet & Roy 2011], and approach that allows systems to improvise in a given style or along with any other musicians.

These later examples, which could probably have also been included under other computer music research areas such as "Machine Listening", do only scratch the infinite potential that *Music Information Research* can bring, in the form of "musical experts", to any type of computer assisted music creation, be it designed for expert musicians and performers or for inexperienced and novel users.

Art Installations

The advent of tangible interfaces, interconnected spatially-tracked objects, open networks and environments, context-driven and location-based applications, and immersive discovery experiences, encourage physical explorations of music. Sonic art installations are being increasingly used to promote

[7] http://www.steinberg.net/en/products/ios_apps/loopmash.html

[8] http://www.freesound.org/

[9] http://mtg.upf.edu/technologies/essentia

commercial products. Tangible interfaces increase the potential of using music information in conjunction with public performance. Open networks and interconnected objects encourage collaborative multi-location music applications. Music Information Research offers increased possibilities of collaboration, sharing, expressive and explorative interaction, immersive installations and ludic explorations in physical space.

Art installations have increasingly been using databases of recorded sounds from field recordings or sounds generated through real-time interaction in order to trigger various behaviours. Installations such as *Sound Mapping London Tea Houses,* exhibited at the Victoria and Albert Museum in 2011 in conjunction with the G-Hack group from Queen Mary, University of London, provide an example of a context for the deployment of Music Information Research in the arts. Within this project the audience can engage with London via a physical map containing audio sound bites from teahouses in various areas across the city. These sound bites can be accessed by moving a sensor-embedded teapot around the map, allowing users to virtually experience the different areas of London through snippets of conversations, which immerse the audience into a range of cultures and every-day activities [G-Hack, 2011]. Other artists, such as Bill Fontana, explore the integration of unfamiliar sounds into new physical environments. Fortana uses the transformative properties of sound to alter the way we experience and understand space and time. In collaboration with The Welcome Collection, Fontana’s recent sound installation *White Sound: An Urban Seascape* (http://www.resoundings.org/) streams the sounds from the seaside in Dorset into a busy London street. It is through the introduction of this unfamiliar virtual rhythm and sound of waves crashing that Fontana transforms the city. Through this action Fontana is able to translate our sense of space and time from the stresses of London to the seaside, where time is readily available and can be enjoyed at our leisure [Fontana, 2011]. The Guardian has also recently used this combination of technical tools and audio sound to skew time and place in order to tell the unconventional narrative of the spatial history within the Kings Cross area of London. By downloading the *London – Streetstories App* and physically exploring the Kings Cross area with headphones, the audience can literally discover the past and hear the stories of places that one may have previously overlooked as a part of the every-day landscape [Panetta, 2012].

Immersive discovery experiences also encourage physical explorations of music. Using ideas of social networking and music recommendation in physical space, *decibel 151* used Binaural Audio to create a spatial audio environment which highlighted relationships between human trajectories and music recordings. It was first presented in the "Information Aesthetics" category at SIGGRAPH 2009 in New Orleans, and used location-related rare recordings of indigenous music from the Southern States. The listeners, encouraged to move around the space others have walked on heard their tracks in surround sound audio. Spatial interpersonal relationships thus evolved between layers of historic recordings and human trajectories. Even though it was conceived as an art installation, decibel 151 shows the impact that Music Information Research-based installations can have on commercial applications. It opened up possibilities for generating online environments for music recommendations on social networking sites, where members could enter a virtual space in order to hear what other participants were listening to, or by taking this concept out onto the street, multiple participants could choose to allow other headphoned individuals to hear what they were listening to and 'hear' people as they walked towards them. In this way participants who have disconnected themselves from the street environment via their headphones are able to reconnect using the same means they had used for disconnection. In this final embodiment, decibel 151 could serve to address the paradox of listening as a means of environmental sound insulation, disconnection of the individual from a crowd, interaction of the isolated individual with other isolated individuals, and thus open up opportunities for their reconnection [Magas et al, 2009].

Play.Orchestra (http://www.milkandtales.com/playorchestra.htm) was an interactive installation on the South Bank. Seats were set up in the positions of an orchestra and passers-by were encouraged to sit down and thus 'play' one of the instruments. Through the random and ludic actions of passers-by the orchestra is played, blurring the borders between audience and player, amateur and professional, free and premium. *Global String* (http://www.ataut.net/site/Global-String) is another installation which requires user engagement in order to produce an effect. Atau Tanaka designed Global String in 1998, in partnership with Kasper Toeplitz, as a way of metaphorically wrapping a musical string around the world. This project explores the idea of communication via non-linguistic but musical interaction and collaboration among people, that can pluck the physical string in one gallery space in order to resonate the string in a sister gallery space across the world. Extending Tanaka's work with Sensorband (http://www.ataut.net/site/Sensorband) Global String presents some of the unique ways to further develop the use of MIR in relation to notions of connectivity and interactivity in contemporary media art [Tanaka, 1998 - present].

Atau Tanaka has been working with the idea of connectivity and networks on a mass scale. However there are also artists who are trying to understand the complex networks and sensory information of the human body. Since around 2006 Daito Manabe has been working on the project Electric Stimulus (http://www.creativeapplications.net/maxmsp/electric-stimulus-maxmsp/) which attempts to understand how sensory information in the human body can be used to create an instrument. By sending a series of targeted shocks to a person's face and outputting the results via the Arduino physical computing platform to generate sound, Manabe literally plays the body as a sensory network for outputting sound and expression through the application of shock waves to particular nerve centres. Using the human body as an instigator in the generation of sound and music is an interesting and growing research area, however it is more commonly thought about in relation to gestural technology. This is a sector of research in which our growing understanding of human engagement and ergonomic interaction is allowing us to generate more intuitive interfaces which work with the natural movements of the human body. Another example of technology engaging with the body to produce sound is the *Serendiptichord dance* (http://www.youtube.com/watch?v=-k5RlIjS-o8). Through the act of performing physically with the circular instrument within the dance there is an odd unity between the production of performance and sound. As the development of tangible interfaces increases its proximity to the body we will continue to see a development in the area of gestural music, a trend that can already be seen in the increasing development of gestural music generating apps (e.g. *Reactable*) [Buehler, 2009].

Music and technology therefore may respond to the body, engage with a space or add layers of sensory information, but music produced may also respond to an environment and be context and location-driven. This can be very clearly seen within the *Variable4* (http://www.variable4.org.uk/about/intro) project. Variable4 is an environmental algorithmic weather machine which creates sound, installed within a particular location where an audience can gather round. Depending on the weather Variable4 generates a unique musical composition which reflects the changing atmosphere of that particular environment. It is easy to comprehend how the addition of a musical composition within a space can have an emotional response [Bulley and Jones, 2012]. However what if we cannot see or sense the factor which is generating the music? This is the case with the *Radioactive Orchestra* (http://www.nuclear.kth.se/radioactiveorchestra/), a group of scientists from the Royal Institute of Technology collaborating with the artist Axel Boman aim to produce a musical sequence from radioactivity. By translating nuclear isotopes to sound frequencies they have allowed people to engage with an intangible force and thus musically experience nuclear physics.

Another potential direction is the combination of MIR techniques for multimodal creation. For example, since September 2011 Barcelona's city council has installed an automatic water and lights choreographies generator for the *Magic Fountain* of Montjuic (one of the main tourist attractions of the city), based on MIR techniques (more concretely on the *Essentia* engine, http://mtg.upf.edu/technologies/essentia). This system allows the person in charge of creating a choreography for the fountain, to pick up a musical mp3 track, decide among several high-level parameters' tendencies (such as the average intensity, contrast, speed of change, the amount of repetition, or the main color tonalities of the desired choreography), and the system generates automatic, music-controlled choreographies at the push of a button [Reactable Systems, 2011].

The use of MIR within the art sector can help with the proposed goal to widen the scope of this research area, ensuring its focus is centered on quality of experience with greater relevance to human networks and communities. As this range of artistic expressions have shown, MIR has a profound impact on the way we as human beings understand space, time and even our own bodies. The arts are uniquely placed, with a degree of freedom from the commercial sector, to have an immense impact on the way we understand MIR and its many applications. It has the ability to experiment with both new technology and concepts alike and push them to their limits. It is this unpressured innovation that will uncover the nuances of how technology is infiltrating our everyday life and uncover the future of how this will develop in the sector of Digital Music.

References

- [Buehler, 2009] Buehler, H. Serendiptichord Dance. Accessed at: http://www.youtube.com/watch?v=-k5RlIjS-o8
- [Bulley and Jones, 2012] J. Bulley, D. Jones. Variable4. Accessed at: http://www.variable4.org.uk/about/intro
- [Fontana, 2011] B. Fontana. White Sound: An Urban Seascape. Accessed at: http://www.resoundings.org/
- [G-Hack, 2011] G-Hack. Sound Mapping London Tea Houses, supported by the National Lottery through Arts Council England, first exhibited as part of Part of Chi-TEK Tea Party by MzTEK at the Victoria and Albert Museum, London during Digital Design Weekend, 24-25 September 2011.
- [Heide et al, 1993-2003] E. Heide, Z. Karkowski, A. Tanaka, A. Sensorband. Accessed at: http://www.ataut.net/site/Sensorband
- [Maestre et al, 2009] E. Maestre, R. Ramírez, S. Kersten. and X. Serra. Expressive Concatenative Synthesis by Reusing Samples from Real Performance Recordings. Computer Music Journal 33 (4): 23–42.

- [Molina et al, 2011] P. Molina, Haro, M., and Jordá, S. BeatJockey: A new tool for enhancing DJ skills. Proceedings of the International Conference on New Interfaces for Musical Expression (NIME) (pp. 288-291). Oslo, Norway.
- [Magas et al, 2009] M. Magas, Stewart, R., and Fields, B. decibel 151: Collaborative Spatial Audio Interactive Environment. In ACM SIGGRAPH.
- [Pachet & Roy 2011] Pachet, F., and Roy, P. (2011). Markov constraints: steerable generation of Markov sequences. Constraints, 16(2), 148-172. Springer Netherlands.
- [Panetta, 2012] F. Panetta. King's Cross, London - Streetstories app for iPhone and Android. Accessed at: http://www.guardian.co.uk/help/insideguardian/2012/mar/21/kings-cross-london-streetstories-app
- [Reactable Systems, 2011] Reactable Systems. Barcelona Magic Fountain at Montjuic. Installed in 2011. http://w3.bcn.es/V01/Serveis/Noticies/V01NoticiesLlistatNoticiesCtl/0,2138,1653_1802_2_1589325361,00.html
- [Schwarz, 2005] D. Schwarz. Current research in Concatenative Sound Synthesis. Proceedings of the International Computer Music Conference (ICMC)
- [Tanaka, 1998 - present] A. Tanaka. Global String. Accessed at: http://www.ataut.net/site/Global-String
- [Višnjić 2010] F. Višnjić. CreativeApplications.Net, Electric Stimulus. Accessed at: http://www.creativeapplications.net/maxmsp/electric-stimulus-maxmsp/
- [Zils and Pachet, 2001] A. Zils, F. Pachet. Musical Mosaicing. Proceedings of the COST G-6 Conference on Digital Audio Effects (DaFx-01), University of Limerick: 39–44, retrieved 2011-04-27

4.3 Research and educational applications

By its nature, MIR is a multi-disciplinary field, and so it is no surprise that MIR outputs have been put to use in research settings outside of music informatics. The most notable impact has been in musicology, where MIR tools have become standard “tools of the trade” for a new generation of empirical musicologists. MIR also shows a lot of promise for educational applications, including music appreciation, instrument learning, theory and ear training, although most existing applications are still at an experimental stage. In this section we examine the relevance of MIR outputs to research and education. We also discuss the benefits and barriers to creating sustainable research outputs - papers, software and data that can be reused to verify or extend published work.

4.3.1 State of the art

A. Research

The use of technology in music research has a long history (e.g. see [Goebl et al., 2008] for a review of measurement techniques in music performance research). Before MIR tools became available, analysis was often performed with hardware or software created for other purposes. For example, Repp used software to display the time-domain audio signal, and he read the onset times from this display, using audio playback of short segments to resolve uncertainties [Repp 1990; 1992]. This methodology required a large amount of human intervention in order to obtain sufficiently accurate data for the study of performance interpretation, limiting the size and number of studies that could be undertaken.

For larger scale studies, automatic analysis techniques are necessary. For example, the beat tracking system BeatRoot [Dixon 2001] has been used in studies of expressive timing [Widmer et al., 2003;

Grachten et al., 2009; Flossmann et al., 2009]. A more general framework for visualisation and annotation of musical recordings is Sonic Visualiser [Cannam et al., 2006; 2010], which has an extensible architecture with analysis algorithms supplied by plug-ins. Such audio analysis systems are becoming part of the standard tools employed by empirical musicologists [Leech-Wilkinson, 2009; Cook, 2004; 2007], although there are still limitations on the aspects of the music that can be reliably extracted, with details such as tone duration, articulation and the use of the pedals on the piano being considered beyond the scope of current algorithms [McAdams et al., 2004].

For analysing musical scores, the Humdrum toolkit [Huron, 1999] has been used extensively. It is based on the UNIX operating system's model of providing a large set of simple tools which can be combined to produce arbitrarily complex operations. Recently, music21 [Cuthbert and Ariza, 2010] has provided a more contemporary toolkit, based on the Python programming language.

B. Education

Education is one of the more understudied and yet promising application domains for MIR. While Piaget's constructivism and Papert's constructionism are classics of pedagogy and interaction design relating to children, mash-up, remix and recycling contents might be considered a much more controversial and radical approach, especially for the social, ethical and legal implications it conveys. However, it is undeniable that young people are embracing remix en masse, and it is integral to how they make things and express ideas. The cultural practices of mash-up and remix brought to school, will force us to rethink the role of teachers as part of this knowledge-building process (Erstad, 2008), and the development of learning strategies that support such models of creation represents an ongoing challenge as it defies the current model of schooling, with students taking a more active role in developing knowledge. The entrance of MIR-powered tools for musical education and creation among younger children opens a new line of research for suitable novel interfaces.

MIR is also seeing uptake in more traditional instrument learning scenarios, via provision of feedback to learners in the absence of a teacher, interactive ear training exercises (Karajan iPhone apps), automatic accompaniment [Dannenberg and Raphael, 2006], page turning [Arzt et al., 2008] and enhanced listening (iNotes: Orchestral Performance Companion).

C. Reproducible Research

Much computational science research is conducted without regard to the long-term sustainability of the outcomes of the research, apart from journal and conference publications. Other outcomes, such as research data and computer software, are stored on local computers, and are lost over time as projects end, students graduate and equipment fails and/or is replaced. Enormous effort is invested in the production of these outputs, which have great potential value for future research, but the benefit of this effort is rarely felt outside of the research group in which it took place. Arguments for sustainability begin with the cost-savings that result from re-use of software and data, but extend to other issues more fundamental to the scientific process. These are enunciated in the "reproducible research" movement [Buckheit and Donoho, 1995; Vandewalle et al., 2009], which promotes the idea that, along with any scientific publication, there should be a simultaneous release of all software and data used in generating the results in the publication, so that results may be verified, comparisons with alternative approaches performed, and algorithms extended, without the significant overhead of reimplementing published work.

Various practical difficulties hinder the creation of long-term sustainable research outputs. The research software development process is usually gradual and exploratory, rather than following standard software engineering principles. This makes code less robust, so that it requires greater effort to maintain and adapt. Researchers have varying levels of coding ability, and may be unwilling to publicise their less-than-perfect efforts. Even when researchers do make code available, their priority is to move on to other research, rather than undertake additional software engineering effort that might make their research more usable. Such software engineering efforts might be difficult to justify in research funding proposals, where funding priority is given to work that is seen to be "research" over "development" efforts. Also, research career progression tends to be awarded on the basis of high-impact papers, while software, data and other outputs are rarely considered. Another perceived difficulty is that public release of software might compromise later opportunities for commercialisation, although various licenses exist which allow both to occur [Stodden, 2009].

To these general problems we may add several issues specific to the music information research community. The release of data is hindered by copyright regulations, particularly relating to audio recordings, but this is also relevant for scores, MIDI files, and other types of data. The laws are complex and vary between countries. Many researchers, being unsure of the legal ramifications of release of data, prefer the safer option of not releasing data. Reliance on specific hardware or software platforms also makes code difficult to maintain in the longer term. One solution for obsolete hardware platforms is the use of software emulation, as addressed by the EU projects PLANETS and KEEP. For music-related research, such general-purpose emulation platforms might not be sufficient to reproduce audio-specific hardware [Pennycook, 2008].

In the MIR community, great effort has been expended to provide a framework for the comparison of music analysis and classification algorithms, via the Music Information Retrieval Evaluation Exchange (MIREX, http://music-ir.org/mirex/wiki/MIREX_HOME), which has been running since 2005. More recently, the Mellon-funded NEMA project (http://nema.lis.illinois.edu/?q=node/12) developed a web service to allow researchers to test their algorithms outside of the annual MIREX cycle. Although there are a growing number of open-access journals and repositories for software and data, there are obstacles such as publication costs and lack of training which hinder widespread adoption. Projects addressing the training aspect are the Sound Software (www.soundsoftware.ac.uk) and the Sound Data Management Training (http://rdm.c4dm.eecs.qmul.ac.uk/category/project/sodamat) projects.

D. Digital library applications

A digital library (DL) is a professionally curated collection of digital resources, which might include audio, video, scores and books, usually accessed remotely via a computer network. Digital libraries provide software services for management and access to their content.

Music Digital Librarians were among the instigators of the ISMIR community, and the first ISMIR conference (2000) had a strong DL focus. Likewise the contributions from the Music IR community to DL conferences (Joint Conference on Digital Libraries, ACM Conference on Digital Libraries, IEEE-CS Conference on Advances in Digital Libraries) were numerous. This could be due to the fact that at the end of 90s, musical libraries moved to digitization of recordings and to multi-information access (video, score images, and text documents such as biographies and reviews) to create multimedia libraries (Fingerhut 99, Dunn 2001, McPherson 2001). In this first trend, the technological aspects of these libraries relied mainly on the server, database, media digitization, text search, and synchronization (often manual) between media. Today this trend still exists and is accessible online for a wide audience. Examples of this are the "Live TV" of the Citè de la Musique (large audience) with synchronizing video concerts with libretto, scores and comments.

A second trend, that appeared in the mid-2000s, reverses the relationship between Libraries and MI Research and Technology. Research and technology enable content estimation, visualization, search and synchronization, which are then used in the context of Digital Libraries to improve the usability and access of the multi-documents in libraries (online or not). Examples of this are: inclusion of automatic audio summaries in the IRCAM Library (Mislin 2005), the Bachotheque to compare automatically synchronized interpretations of a same piece (Soulez 2003), optical score recognition and audio alignment for the Bavarian State Library (Damm et al., 2011). Also, thanks to the development of technologies (Flash, html5, Java-Script), the de-serialization of media becomes a major theme, along with improved browsing and access to the temporal aspect of media. New concepts of interfaces to improve the listening have been developed which make use of time-based musical annotations (Ecoute augmentee/Increased-listening, or today's SoundCloud).

A third trend concerns the aggregation of content and the use of user-generated annotation. The content of dedicated libraries can be aggregated to form meta-libraries (e.g. www.musiquecontemporaine.fr) using the shared protocol OAI-PMH. Content can be distributed over the web or aggregated to a local collection. Using Semantic Web technologies such as Linked Data and ontologies, web content can be re-purposed (e.g. BBC's use of the Music Ontology). This trend is also found in the new form of music access (such as Spotify) which aggregates content related to the music item (AMG reviews, wikipedia artist biography).

Comparing the suggestions of [Bonardi, 2000] and the observations of [Barthet and Dixon, 2011] a decade later, it is clear that much work is still to be done before MIR technology is fully incorporated into traditional libraries. The European ASSETS project (http://www.assets4europeana.eu/), working with the Europeana multi-lingual European cultural collection, aims to improve search and browsing access to the collection, including multimedia objects.

References

- [Arzt et al., 2008] A. Arzt, G. Widmer, S. Dixon (2008). Automatic Page Turning for Musicians via Real-Time Machine Listening. Proceedings of the 18th European Conference on Artificial Intelligence, pp 241-245.
- [Bonardi, 2000] A. Bonardi (2000). IR for Contemporary Music: What the Musicologist Needs. Proceedings of the International Symposium on Music Information Retrieval.
- [Barthet and Dixon, 2011] M. Barthet and S. Dixon (2011). Ethnographic Observations of Musicologists at the British Library: Implications for Music Information Retrieval. 12th International Society for Music Information Retrieval Conference, pp 353-358.
- [Buckheit and Donoho, 1995] J. B. Buckheit and D. L. Donoho (1995). WaveLab and reproducible research. Department of Statistics, Stanford University, Technical Report 474.

- [Cook, 2004] N. Cook (2004). Computational and Comparative Musicology, in Empirical Musicology: Aims, Methods, and Prospects, Ed. E. Clarke and N. Cook, pp. 103-126, Oxford University Press: New York.
- [Cook, 2007] N. Cook (2007). Performance Analysis and Chopin's Mazurkas, Musicae Scientae, 11 (2), 183-205.
- [Crawford and Gibson, 2009] T. Crawford and L. Gibson (Ed.) (2009). Modern Methods for Musicology: Prospects, Proposals, and Realities. Digital research in the arts and humanities, Ashgate.
- [Cuthbert and Ariza, 2010] M.S. Cuthbert and C. Ariza (2010). music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data. 11th International Society for Music Information Retrieval Conference, 637-642.
- [Damm et al., 2011] D. Damm, C. Fremerey, V. Thomas, and M. Clausen. A demonstration of the Probado music system. In Proc. of ISMIR, 2011.
- [Dannenberg and Raphael, 2006] R.B. Dannenberg and C. Raphael (2006). Music score alignment and computer accompaniment. Communications of the ACM, 49(8), 38-43.
- [Dixon et al., 2002] S. Dixon, W. Goebl and G. Widmer (2002). The Performance Worm: Real time visualisation based on Langner's representation. Proceedings of the 2002 International Computer Music Conference, Göteborg, Sweden, 361–364.
- [Dunn and Isaacson, 2001] J. W. Dunn and E. J. Isaacson. Indiana University digital music library project. In Proc. of JCDL, 2001.
- [Erstad, 2008] Erstad, O. (2008). Trajectories of remixing: Digital literacies, media production, and schooling. Digital Literacies, 177–202.
- [Fingerhut 1999] M. Fingerhut. The IRCAM multimedia library: a digital music library. In IEEE Forum on Research and Technology Advances in Digital Libraries (IEEE ADL'99), Baltimore, MD (USA), 1999.
- [Flossmann et al., 2009] Flossmann, S., Goebl, W., & Widmer, G. (2009). Maintaining skill across the life span: Magaloff's entire Chopin at age 77. Proceedings of the International Symposium on Performance Science, pp. 119–124.
- [Grachten et al., 2009] Grachten, M., Goebl, W., Flossmann, S., & Widmer, G. (2009). Phase--plane representation and visualization of gestural structure in expressive timing. Journal of New Music Research, 38(2), 183–195.
- [Huron, 1999] Huron, D. (1999). Music Research Using Humdrum: A User's Guide. Stanford, California: Center for Computer Assisted Research in the Humanities, 414 pages.
- [Leech-Wilkinson, 2009] D. Leech-Wilkinson (2009). The Changing Sound of Music: Approaches to Studying Recorded Musical Performance, CHARM: London, www.charm.kcl.ac.uk/studies/chapters/intro.html (Accessed 11/04/2012).
- [McAdams et al., 2004] S. McAdams and P. Depalle and E. Clarke (2004). Analyzing Musical Sound, in Empirical Musicology: Aims, Methods, and Prospects, Ed. E. Clarke and N. Cook, pp. 157-196, Oxford University Press: New York.
- [McPherson and Bainbridge, 2001] J. R. McPherson and D. Bainbridge. Usage of the meldex digital music library. In Proc. of ISMIR, 2001.
- [Mislin and Peeters, 2005] F. Mislin and G. Peeters. Automatisation de la production et de la mise en ligne de resumes sonores. Master's thesis, ISTY, 2005.

- [Pennycook, 2008] B. Pennycook (2008). Who Will Turn the Knobs when I Die? Organised Sound, 13 (3), 199-208.
- [Raimond et al., 2009] Y. Raimond, C. Sutton, and M. Sandler. Interlinking music-related data on the web. IEEE Multimedia, 16(2):52–63, April-June 2009.
- [Repp, 1990] B.H. Repp (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. Journal of the Acoustical Society of America, 88(2), 622-641.
- [Repp, 1992] B.H. Repp (1992). Diversity and commonality in music performance: An analysis of tiiming microstructure in Schumann's "Träumerei". Journal of the Acoustical Society of America, 92(5), 2546-2568.
- [Stodden, 2009] V. Stodden (2009). The legal framework for reproducible scientific research: Licensing and copyright. Comput. Sci. Eng. 11, 35–40.
- [Soulez et al., 2003] F. Soulez, X. Rodet, D. Schwarz, et al. Improving polyphonic and poly-instrumental music to score alignment. In Proc. of ISMIR, Baltimore, Maryland, USA, 2003
- [Vandewalle et al., 2009] P. Vandewalle, J. Kovacevic and M. Vetterli (2009). Reproducible Research in Signal Processing - What, why, and how. IEEE Signal Processing Magazine, Vol. 26, Nr. 3, pp. 37-47.
- [Widmer et al., 2003] G. Widmer, S. Dixon, W. Goebl, E. Pampalk and A. Tobudic (2003). In Search of the Horowitz Factor. AI Magazine, 24 (3), 111-130.
- [Widmer et al., 2008] G. Widmer, S. Dixon, P. Knees, E. Pampalk and T. Pohle (2008). From Sound to Sense via Feature Extraction and Machine Learning: Deriving High-Level Descriptors for Characterising Music, in Sound to Sense - Sense to Sound: A State of the Art in Sound and Music Computing, Ed. P. Polotti and D. Rocchesso, pp. 161-194.

4.4 Creative industries applications

Music Information Research is becoming increasingly relevant for creative and commercial installations, applications and environments. The potential for using music information is in conjunction with creative marketing tools, mobile apps, gaming, commercial installations, environmental installations, indoor and outdoor events.

4.4.1 State of the art

The exceptional growth of the Digital Music Market in recent years is stated within IFPI’s (International Federation of the Phonographic Industry) annual Digital Music Report [IFPI, 2012] we have seen a remarkable growth of 8 per cent globally within the industry. Digital channels now account for an estimated 32 per cent of record company revenues globally, up from 29 per cent in 2010 (http://www.ifpi.org/content/library/DMR2012.pdf). It is this financial and user growth within this sector that creates a new space for entrepreneurs and creatives to consider new business models and applications for continuing research into MIR.

Since the advent of the Apple App Store (launched mid 2008) there has been a surge of music service applications developed and popularized. Shazam and Pandora were two of the early innovators, who recognised that the Future of Music was in music information. Both launched successful services and were listed in the top five music applications by NPR Music in 2008 (http://www.npr.org/blogs/allsongs/2008/07/top_5_iphone_music_application_1.html). These services worked by creating useful social information through retrieval algorithms or music descriptors which allowed people to experience music in a new way.

From music production to the multitude of ways we consume and interact with music, acquiring a taste for a particular artist or genre is being drastically changed. In fact, the entire basis of the music industry is going through a major overhaul. As stated in section *4.1 Search and Discovery Applications* music is being produced and published at an excessive rate. In this new environment of overwhelming choice how do commercial applications and recording studios develop strategies, which take into account the nature of the internet and thus a growing user demand for a more intimate relationship with artists and recording studios? This new level of involvement and choice on the part of users is currently being explored (using recent developments in MIR) largely through a greater level of engagement with social media. For example the Coke Music 24 hr challenge (http://www.nexusinteractivearts.com/work/hellicar-lewis/coke-music-24hr-session-with-maroon-5) allowed fans of Maroon 5 to contribute to and to support the development of the band's newest single within a live 24hr period via their social networks. Social recommendation is another way that social media is being used to bring the wants and needs of fans closer to the music industry. There are now several companies who base their entire business model on the idea of social recommendation and access on demand. This means that users can move through their favourite tracks and recommendations in a fluid manner without having to buy each individual album/ track they are interested in. As one of the current market leaders Spotify has enjoyed an economic growth of 1 million paying users in March 2011 to 3 million paying users by January 2012 (http://www.guardian.co.uk/media/2012/jan/29/spotify-facebook-partnership-apps). This exponential growth can be majorly attributed to the company's recent integration with Facebook, which has created an inter-linked network of social music recommendations within groups of friends. The application Serendip (http://serendip.me/) also uses this technique. Via a seamless Twitter integration it creates a real time social music radio allowing users the opportunity to independently choose 'DJs' from their followers and share songs across a range of social media.

The industry of music service and music making applications has also been growing steadily over the last four years, offering ever more tailored information and exciting opportunities for MIR. As these services have been developing, so has the hardware. Smart music players are now embedded in many phones and devices with features like geo-location, touch screens, mobile internet access and movement sensors. Application developers are using this new range of sensory information to create more immersive sonic experiences and music generators. One recent example of this is the Musicity project (http://musicity.info/home/) which has been nominated in the London Design Museum "Design of 2012" awards. This location-based application inspires a new exploration of the city by encouraging people to visit new interesting locations in order to collect and experience music written specifically for that space. The fact that the audience have to visit a specific area in order to download the music has a positive effect on memory and impact. In the process of physically finding their music it creates a new relationship with both the artist and place they have just 'discovered' and means people are more likely to want to discover more. This approach has also been used in reverse where the environment is used to generate music. London-based RjDj (http://rjdj.me/) creates a mash-up of music by embedding sounds it picks up within your environment to create a continuous music track. There are also several apps which allow smart devices, for example the iPhone, to be transformed into portable musical instruments which engage with the body and allow for spontaneous performances. Using the latest technologies in human computer interaction, music technology and graphics, the Reactable app (http://www.reactable.com/products/mobile/: adapted from the larger physical "Reactable") allows users, ranging from amateur to professional, to improvise music in an intuitive and visual way by moving virtual objects around a touchscreen which changes the tone, pitch, tempo etc of the musical piece and thus creates unique opportunities for creative play and performance. Bloom (http://www.generativemusic.com/) has a similar function to Reactable, however this app relies on a visual output which is created by translating sonic waves into graphical interfaces. This app is described as part instrument, part composition and part artwork and is another example of the new and unique meshings that research into MIR can create, adapting the way we understand performance, instrument, art, play and environment.

Use of fun (e.g. Volkswagen's Fun Theory, http://www.thefuntheory.com/) and gaming is another developing application for MIR with various research and commercial possibilities. The musical interaction team at IRCAM has been working with motion sensors embedded within a ball to explore some of these concepts of integrating fun, gaming and musical experience. The Urban Musical Game (http://www.mires.cc/?q=node/105) breaks down some of the boundaries between audience and musician by producing a sound environment through the introduction of a musical ball. By throwing the ball different sounds are produced, thus becoming an instrument as a by-product of having fun and just being involved in the game. Joust (http://gutefabrik.com/joust.html) is another interesting example of how our relationship with music and environment is changing. Joust is a spatial musical gaming system in which two players are given motion controllers and have to circle each other until the music speeds up. They then have a brief window to tag their opponent and win the game. The use of rhythm and pace as the instigator of the action is a unique experiential concept which allows for the build up of atmosphere and adrenaline (a device often used in film). The use of music and the body as an integral part of the game immerses people almost instantaneously and allows for an exciting and fun experience. Even at its prototype stage Joust has already been a great success, winning several awards (including the Innovation Award: Game Developers Choice Award 2012).

Moving beyond gaming, the creation of fun musical environments' can also be used as a way of making commercial products memorable and fun. Wrigleys Augmented Reality Music Mixer is one example of this. By using current innovations in augmented reality, communications agency Exposure (http://exposure.net) worked with technical partner Boffswana (http://boffswana.com/) to promote the launch of Wrigleys new range of gum (5Gum) in France (http://5gum.fr/). By simply printing off five distinct symbols and turning on the user's webcam, this website allows users to play the five music genres (relating to the five new gum flavours) via moving and placing their hands over the printed symbols, viewing the virtual reaction to their touch. After they have finished creating their unique music-mix they can then share their creation within their social networks thus further promoting the brand. Another promotional event using immersive musical environments was The Echo Temple at Virgin Mobile FreeFest (http://great-ads.blogspot.co.uk/2011/09/interactive-sound-installation-for.html). This temple installation created a shared experience of making music through the use of motion tracking cameras and fans branded with special symbols. This allowed people throughout the festival to play with this musical installation, bonding with others at the event and creating a more lasting communal memory.

In the Digital Music sector business models are constantly developing to fit the shifting dynamic of new kinds of services. Some companies, such as Spotify, are using an ad-based free service combined with a tier-based subscription where power users pay extra in order to have greater privileges. Other services are selling technologically innovative apps (e.g. RjDj) or collaborating with commercial brands and technology providers (e.g. The Echo Temple). By gamifying their services and producing interactive experiences, innovative companies are working to increase their products' core values. Digital tools have the immense potential to create fantastic opportunities for artists, fans and labels alike. However in this fast-paced development some of the industry is being left behind and it is only through the allegiance of R&D developers, record labels and artists that Europe will be able to competitively move forward in the expanding field of Digital Music.

As this industry develops there will undoubtedly be a call for more engaging uses and applications of MIR. There will also be a greater pressure on artists, recording studios and brands to engage further with the wider community and to work towards having a two-way relationship with their fans and users. As Mark Mulligan states in his 2011 report "digital and social tools have already transformed the artist-fan relationship, but even greater change is coming…the scene is set for the Mass Customization of music, heralding in the era of Agile Music" [Mulligan, 2011]. Agile Music is a framework for understanding how artist creativity, industry business models and music products must all undergo a programme of radical, transformational change. This change must occur at all levels of the industry if it is to be successful in giving users a greater experiential value. By building better user relationships through new technological and social models the music industry will be in a better position to compete with the current mass phenomena of free downloadable music. This potential market could have an immense impact in the creation of new jobs for innovative creative people in Europe.

References

- [Dredge, 2012] S. Dredge. Spotify says Facebook partnership and new apps should allay growth fears. The Guardian. Accessed at: http://www.guardian.co.uk/media/2012/jan/29/spotify-facebook-partnership-apps
- [IFIP, 2012] IFIP. Digital Music Report: Expanding Choice. Going Global. Accessed at: http://www.ifpi.org/content/library/DMR2012.pdf
- [Hilton, 2008] R. Hilton, R. Top 5 iPhone Music Applications. All Songs Considered: The Blog. NPR Music. Accessed at: http://www.npr.org/blogs/allsongs/2008/07/top_5_iphone_music_application_1.html
- [Hellicar & Lewis. 2011] Hellicar & Lewis. Nexus Productions. Coke 24hr session with Maroon 5. Accessed at: http://www.nexusinteractivearts.com/work/hellicar-lewis/coke-music-24hr-session-with-maroon-5
- [Mulligan, 2011] M. Mulligan. Music Formats and Artist Creativity In The Age of Media Mass Customization. A Music Industry Blog Report. Accessed at: http://musicindustryblog.wordpress.com/free-reports/
- [Sylvia, 2011] G. Sylvia. Interactive Sound Installation for the Virgin Mobile FreeFest "The Echo Temple"" GreatAds. Accessed at: http://great-ads.blogspot.co.uk/2011/09/interactive-sound-installation-for.html

E CONCLUSION

This deliverable provided the current status of the review of MIR state-of-the-art conducted by the MIReS consortium.

This document serves as input to Work Package 3 and in particular to D3.3.

The state-of-the-art will be published on the project Wiki for facilitating contributions by the MIR community.

An updated version of this state-of-the-art (including potential contributions by the community) will be an integral part of the Roadmap.