New Frontiers in Music Information Processing (MIP-Frontiers)

Grant Agreement Number: 765068

- Title: State of the art, challenges and potential of knowledge-driven approaches in MIR
- Lead Beneficiary: JKU (Johannes Kepler University)
- Nature: Report
- Dissemination level: Public

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

Outline

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

1. Summary

MIP-Frontiers is a project that focuses on training PhD students (or Early Stage Researchers, ESRs) in the field of Music Information Retrieval (MIR), preparing the next generation of MIR researchers. In doing so, MIP-Frontiers needs to address the main challenges that face the field of MIR. In a recent strategic document (MIReS Roadmap, EU FP7 programme project), these challenges were identified as relating to: data-driven aspects, knowledge driven aspects, and user-driven aspects. Regarding knowledge-driven aspects – the focus of the present report – the main message is that MIR research needs to investigate ways to include and exploit additional knowledge and information into its methods and systems, in order to obtain better and more robust solutions and provide for a more natural communication between system and human user.

The preparation of this report started with an internal discussion, within the consortium, on what knowledge-driven methods can mean for MIR in general, and for the PhD students' projects in particular. It was obvious that each of the 15 MIP-Frontiers projects targets different aspects of knowledge-driven methods, and therefore each has to present and discuss the state of the art, challenges and potentials in its own specific context; and that the report should reflect that.

Thus, each project has presented, and contributed to this report, its specific view of how general musical and extra-musical knowledge can be of help in solving complex music processing tasks. This report has already been, and will be, a valuable resource, demonstrating to the ESRs and to the MIR community the importance of knowledge-driven approaches to MIR research.

2. State of the project

MIP-Frontiers is a four-year project. It started in April 2018, is now at month 18, and all ESRs have been enrolled for at least 9 months. The fellows have all presented their thesis Stage 1. This stage involves learning about the project requirements and mapping out personalized goals and challenges; it also offers a framework for thinking about the possible paths of their research. They need to understand the state of the art, challenges, and approaches of the MIR field.

Although the EU FP7 project MIReS defined the principal challenges in its project Roadmap, the evolution of the field and the actual implication on the ESR projects needed an interpretation and a revision to adapt it to the student context. At the project board meeting held in Barcelona in May 2019, it was discussed how to write and address this report on "State of the art, challenges and potential of knowledge-driven approaches in MIR", especially with a view to how it would be useful for the ESRs and other future MIR researchers.

The report starts with an introduction that describes how the term "knowledge-driven methods and approaches" is understood in the framework of MIP-Frontiers. As the MIP-Frontiers Training Network comprises 15 individual ESRs with different topics in the MIR field, we obtain fifteen different views on relevant scientific background, challenges, opportunities, and solutions.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

3. Introduction

In the project proposal, the MIP-Frontiers consortium identified three big challenges – and thus, from a scientific point of view, research opportunities – for the further development of the MIR field. One of these is the need for more high-level knowledge in MIR systems, and in the research and development process (in addition to large amounts of data). Specifically, this was argued in the proposal as follows: *It is our conviction that musical knowledge will be key to developing the next generation of music processing technologies. While current advances in Deep Learning may seem to imply that almost any recognition task can be learned from (huge amounts of) data alone, we believe that music, as an extremely rich and multi-faceted (physical, physiological, psychological, socio-cultural) phenomenon, will benefit strongly from the inclusion and exploitation of additional (in some cases, extra-musical) knowledge and information. Not only will this lead to musically better solutions, but it will also support more natural communication between system and user, and in this way improve the acceptance of practical music systems (e.g., in search and recommendation). A special aspect of this is multi-modality - the integration of several data channels (video, textual information, ...) in music-centred systems, as an additional source of context and information.*

The objectives of knowledge-driven approaches in MIR, as defined in the project web, https://mip-frontiers.eu/, and in the project proposal, are to research the use of high-level musical knowledge and contextual information for solving hard music processing and recognition tasks; to investigate systematic ways of integrating expert or contextual knowledge into (deep) learning processes; and to demonstrate the beneficial effects of high-level knowledge.

The purpose of the present report is to provide a starting point for this work, by documenting the state of the art, current challenges, and corresponding potential of knowledge-driven research approaches to MIR problems. As "knowledge-driven methods in MIR" as a general concept is far too broad for us to be able to give a comprehensive overview, this report will focus on knowledge-related topics as they emerge in the specific research projects (PhD theses) tackled in MIP-Frontiers. To this end, the next section (Section 4) will give an overview of which of these projects are related to knowledge-driven approaches, and in what specific ways. The remainder of the report will then present a discussion of the state of the art, challenges, and potential from the viewpoint of these particular projects or tasks. Section 5 will thus be structured by individual PhD projects. Further details are found in the students' Stage 1 reports.

4. Knowledge-driven methods in MIP-Frontiers

4.1 Musical and extra-musical “knowledge”

“Knowledge” is a complex concept. In the context of the present project, and specifically for the purposed of the project, we interpret it rather loosely as any kind of information that is provided to a (learning) system in addition to raw training data. We will speak of “high-level” knowledge and distinguish this from “low-level” data and features, to emphasise the view that – especially in the context of machine learning – general or domain-specific knowledge can provide constraints and guidance to a system beyond what the raw training data can, because it relates to more abstract concepts and relationships that are not easily inferable from the given data. Exploiting such knowledge thus introduces a top-down aspect into the system.

[1] This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

In the context of MIR, relevant knowledge may come from many different fields, and find its way into music systems in various ways, for example: knowledge about music perception encoded in advanced hand-crafted features; physical/acoustical models of instruments as a source of bias in audio separation or transcription; musicological-stylistic knowledge used as constraints on permitted solutions; or information from other modalities and information sources (video, user and performance context, etc.) that provides additional guidance and context to a music processing system.

The rest of this section lists those projects in MIP-Frontiers where knowledge-related aspects as defined above play a role, in one way or another, and briefly explains why. For each of these projects, Section 5 will then describe the current state of the art and corresponding challenges and opportunities.

4.2 MIP-Frontiers projects with knowledge-driven aspects

QMUL1: "Representation Learning in Singing Voice"

The goal of this project is to develop deep learning based systems to solve singing voice related problems that are commonly tackled in the music software industry. This project focuses on a few of these problems including lyrics/phoneme transcription, audio-to-lyrics alignment, and vocal technique identification. The research exploits musical knowledge to define target information to extract from singing voice signals, and design learning algorithms based on these definitions. Once extracted, new tools will be developed to represent this information in music notation and for the analysis of the singing voice.

QMUL2: "Improving Polyphonic Transcription through Instrument Recognition and Source Separation"

Project QMUL2 focuses on better understanding the aspects and qualities of music sounds that are related to the timbre of musical notes and that force us to represent them differently in the staff notation. The specific research goal is to be able to associate each sound to the correct instrument, as well as detect and recognise different playing techniques (pizzicato, legato and vibrato, for example) used throughout the music by the same instrument, so that proper symbols can be applied in the transcription to represent them. This can be done, for example, by including knowledge from the field of musical acoustics (detailed models of the mechanics of instruments, and the resulting range of sounds they can produce) and also from the field of music perception (characteristics of sounds that affect our perception of timbre, such as the relation between the energy of the harmonics that each sound stimulates) in the system, which allows us to create more powerful and efficient networks.

QMUL3: "Leveraging User Interaction to Learn Performance Tracking Music Alignment"

The project aims at providing a way to navigate among various music representations in a unified manner, lending itself applicable to a myriad of domains like music education, performance, enhanced listening, automatic accompaniment and so on. This project (QMUL3) has a moderate knowledge-driven component, and it is combined with more dominant data-driven and user-driven components.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

QMUL4: "Robust Timbre Analysis for Query by Vocal Imitation"

While this project relies on deep learning algorithms to have the final say when linking vocal imitations with their original sounds, traditional knowledge of timbre is key to allow these methods to achieve the highest performance. This is due to the fact that there is not enough vocal imitation data available for these algorithms to be self-sufficient, and previous knowledge about the task becomes an essential input for these to consider.

QMUL5: "Adversarial attacks to understand deep learning models for music"

This project aims to bridge the gap between deep learning approaches to MIR and traditional knowledge based approaches where we understood what features were used to perform a task.

UPF1: "Facilitating Interactive Music Exploration"

Project UPF1 aims to utilize different semantic categories of tags, such as genre, mood/theme, and context to complement each other to cover different aspects of the music exploration space. The semantics of the user's goals will be used to identify the relevant areas of the constructed exploration space.

UPF2: "Methods for Supporting Electronic Music Production with Large-Scale Sound Databases"

The goal of project UPF2 is to develop novel methods for browsing loops in large collections of sounds. In order to better characterise the loops in these collections, we can employ algorithms which use knowledge derived from music theory and perception.

UPF3: "Identifying and Understanding Versions of Songs with Computational Approaches"

This project aims to build version identification systems that would provide both a new notion of music similarity from a Music Information Retrieval perspective and a practical tool for music monitoring services from an industrial perspective. Our goal is to use computational methods that would aid us gaining a more comprehensive understanding of relations among versions in order to develop more insightful and scalable systems. We will put a particular focus on the importance of domain knowledge in solving the version identification task.

TPT1: "Behavioural Music Data Analytics"

Project TPT1 aims at improving music recommendation systems by integrating contextual information about a user in the recommendation process. A user's contextual information is defined as the external factors that affect the user's music preferences at any given time. For example, the user's activity, location, or time of the day are considered as contextual information that changes user's preferences. For certain activities the user would prefer to listen to energetic music while in some others he/she would prefer to listen to calming music. Hence, this external information is considered the knowledge-driven aspect of this project. The aim of the project is to rely on the raw audio data (data-driven) plus the contextual information (knowledge-driven) to provide recommendations that would suit the user's taste and current context.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

TPT2: "Voice Models for Lead Vocal Extraction and Lyrics Alignment"

Audio source separation is the task of extracting individual sound sources from a mixture. Project TPT2 aims at developing robust audio source separation methods for singing voice extraction. The specific research goal is to exploit available prior knowledge about the singing voice in the form of lyrics transcripts, which are often widely available in the internet and can otherwise be produced by users without any musical knowledge required. The sounds produced by a singer can be seen as a combination of pitch (determined by the vocal cords) and pronunciation of words (determined by the vocal tract, position of tongue, jaw, etc.). While no information about the pitch is contained in the lyrics transcripts, they contain information about pronunciation-related aspects of singing voice. Specifically, they indicate which utterances will appear in which order in the singing voice. Project TPT2 researches possibilities to integrate this knowledge into state of the art data-driven singing voice separation methods and investigates if this additional knowledge can lead to improved separation performance.

TPT3: "Multimodal Movie Music Track Remastering"

Project TPT3 will heavily exploit multimodal information to improve automatic movie music track remastering. The specific research goal is to exploit prior information about the musical sources to enhance the separation process. This is done by extracting additional information of the attended source from the neural response of the user who is focusing on a given instrument while listening to polyphonic music mixtures.

TPT4: "Context-driven Music Transformation"

The aim of project TPT4 is to enable transforming music in terms of artistic style. Due to the open-ended nature of the task, a fully data-driven solution might be impossible, and some amount of domain knowledge will probably be needed in order to fully define the task and to provide some form of supervision, guiding the algorithm to a meaningful solution. This could be done, for example, by finding and aligning similar music segments in different styles, or by generating a completely synthetic training dataset which will be aligned by design.

TPT5: "Conditional Generation of Audio Using Neural Networks and its Application to Music Production"

The general research goal of project TPT5 is to synthesize audio using conditional Deep Generative Neural Networks and explore applications to music production. Concretely, we consider the use of Generative Adversarial Networks (GANs) to synthesize some musical audio content given prior descriptive information (e.g., pitch, instrument), and some audio representation of pre-existing music content to which the synthesized audio will be adapted.

JKU1: "Large-scale Multi-modal Music Search and Retrieval without Symbolic Representations"

Project JKU1 aims at developing new algorithms for the automatic structuring and cross-linking of large multi-modal music collections, with a focus on audio recordings and sheet music images (acoustic and visual domains, respectively). In addition to massive amounts of musical data (audios, scores in various representations), solving this will require the use of additional higher-level knowledge, at various levels – for instance, knowledge about typical musical section structure; expectations about temporal relations in music; or general knowledge about different musical styles and notation or performance conventions.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

JKU2: "Live Tracking and Synchronisation of Complex Musical Works via Multi-modal Analysis"

Project JKU2 focuses on multi-modality as a source of additional information to guide the hard task of tracking complex musical stage works (operas). In particular, we will need to develop new methods to integrate different extra-musical knowledge sources (e.g., singing voice and speech detection, acoustic event detection, video (when available)) into audio-based music score following algorithms in order to achieve robustness.

5. State of the art, challenges, and opportunities (by project)

5.1 QMUL1 "Representation Learning in Singing Voice"

The ***goal*** of this project is to develop deep learning based systems to solve singing voice related problems that are commonly tackled in the music software industry. This project focuses on a few of these problems including lyrics/phoneme transcription, audio-to-lyrics alignment, and vocal technique identification. The research exploits musical knowledge to define target information to extract from singing voice signals, and design learning algorithms based on these definitions. Once extracted, new tools will be developed to represent this information in music notation and for the analysis of the singing voice.

The ***state of the art*** that we can start from and build upon is as follows. The project focuses on two aspects of singing voice: the verbal content and the vocal techniques. For the latter, we begin with a set of techniques included in VocalSet [1]. More explicitly, the list of vocal techniques contains vibrato, trill, trillo, vocal fry, breathy, belt, inhaled and lip trill. The second focus of this research is the verbal content in singing voice, more specifically the transcription of phonemes and lyrics. In recent years, various versions of Smule – DAMP singing dataset [2] has been released and made available for research. This dataset consists of real world acapella recordings of karaoke songs, performed by various users. Prior research worked on creating clean and aligned subsets of this dataset [3,4] and reported preliminary results for the transcription accuracy using a common DNN-HMM topology. The recent work of Dabike [5] uses a new version of the Smule – DAMP dataset, where prompt-level time annotations are provided, to train a Time-Delay Neural Network (TDNN) and achieved less than 20% word error rate (WER). To this date, that work is the stateof-the-art system for lyrics transcription.

In the recent years of Deep Learning research, a new type of neural network has gained high popularity, which is called the "attention mechanism in neural networks" [6], due to providing us with the ability to visualize and interpret how and what the network is learning. In our research, we will investigate ways to include this mechanism in deep learning architectures designed for lyrics extraction and vocal technique identification tasks.

From the above, we derive a number of ***challenges and opportunities*** for original research. When listening to the singing voice, our perception pays attention to different modes of information depending on the target. For instance, timbral aspects of the voice are mostly determined by its spectral characteristics, meaning that the human auditory system pays more attention to certain frequency bands. On the other hand, understanding the verbal content or the lyrics have temporal dependency to the previously pronounced words and phonemes. Moreover, the characteristics of the pitch variation in time is one of the factors that help vocal experts to distinguish between different vocal techniques. The main challenge of this research

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

is then to design a task-specific neural network architecture with the purpose of learning the aforementioned characteristics of the singing voice and with the purpose of modelling human perception. In addition, due to the lack of a universal ontology for vocal techniques and characteristics, our research will focus on aspects of the singing voice limited by the availability of the open source data.

5.2 QMUL2: "Improving Polyphonic Transcription through Instrument Recognition and Source Separation"

This project explores the interdependencies between instrument recognition and source separation with the final goal of improving polyphonic instrument transcription. The project proposes the creation of a system capable of automatically learning timbre-related features for identifying the different sound types that are being played and capable of separating the music signal into multiple sources based on the detected timbres.

The relevant ***state of the art*** is as follows. Regarding instrument recognition, most of the knowledge-driven methods focus on using handcrafted features to classify the instruments using statistical rules. For instance, [1] uses Mel-Frequency Cepstrum Coefficients (MFCCs) along with Principal Component Analysis (PCA) to reduce dimensionality and Gaussian mixture models (GMM) for classifying solo phrases of 5 instruments. On the other hand, [2] utilises a Modified Group Delay Feature (MODGDF), which is a combination of MFCCs with phase information, to improve the instrument recognition.

Regarding source separation, most knowledge-driven methods are signal processing algorithms built to explicitly model characteristics of the structure of a music signal or to exploit particular spectro-temporal features of musical instruments [3]. In order to separate the melodic voice from the instrumental accompaniment of a music, for example, a useful assumption is to consider the sounds of accompaniment instrumental sources as repetitive patterns in the mixture's spectrogram and the melody line as the non-repetitive part [4].

Furthermore, there are more complex and hybrid methods, where pitch estimation techniques are also included in the separation algorithm. For example, [5] uses a multi-pitch estimation algorithm to identify the pitch contour of the singing voice in order to extract it from the music recording, while [6] mixes analysis in the frequency domain using Independent Component Analysis (ICA) and a method denoted by the author as Amplitude Discrimination with the posterior time domain analysis and pitch estimation to separate singing voice.

There are a number of open ***challenges and opportunities*** for original research. Finding a mathematical set of rules using statistical signal processing techniques in order to perform instrument recognition when there are multiple instruments interacting in a polyphonic scenario is highly complex and impractical. Moreover, the problem of separating an audio signal with multiple harmonic and percussive instruments into multiple sources is too challenging to be tackled by an exclusively knowledge-driven method, as the patterns of spectral characteristics are too difficult to define explicitly.

Therefore, project QMUL2 sees an opportunity of using knowledge-driven approaches to enhance the performance of user-driven methods. In other words, by using domain-specific knowledge, it is possible to create more efficient and powerful machine learning algorithms. For instance, the first contribution of the project was the proposal of a novel convolutional

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

network for performing harmonic-percussive source separation much more efficiently with a greatly reduced number of parameters [7] . This was done by exploiting the fact that percussive signals form vertical patterns in the music spectrogram while harmonic signals tend to form horizontal structures. So, we proposed to use filters (kernels) of vertical and horizontal shapes in the convolutional layers to make the network learn the different time-frequency patterns more efficiently.

5.3 QMUL3 "Leveraging User Interaction to Learn Performance Tracking"

Project QMUL3 is aimed at developing robust music alignment techniques. These will mainly focus on deep learning methods; however, we plan to employ domain knowledge in order to build alignment models which can perform well specifically for the music domain. To this end, we would be employing domain adaptation as well as data augmentation techniques.

Regarding the ***state of the art***, most of the knowledge-based approaches for music alignment are based on Dynamic Time Warping. Originally proposed by [1], dynamic time warping (DTW) is a general method for finding an optimal alignment between two time series by comparing the features of the feature sequences using a local cost function, at each point, with the goal of minimizing the overall cost. The path which yields this minimum overall cost is then the optimal alignment between the two sequences. We briefly mention the works on alignment based on DTW below.

One of the earliest works in music alignment [2] proposes a method for aligning polyphonic audio recordings by first mapping MIDI data to corresponding audio features and then matching these features to the recording, as opposed to symbolic ones. [3] is one of the early works which focus on addressing the space and time complexity of standard DTW-based algorithms for alignment of complex, polyphonic piano music. [4] originally proposed a DTW-based method for online music alignment. It differs from standard DTW-based approaches in the way it computes the alignment. The online alignment is computed incrementally and has a linear time and space complexity.

There have been multiple approaches building on such a standard DTW-based alignment algorithm, such as [5] and [6]. [7] developed a method for automatic extraction of tempo curves from expressive music recordings by comparing performances with neutral reference representations. [8] works on addressing issues dealt in score following by proposing a multilevel matching and tracking algorithm which continually updates and evaluates multiple high-level hypotheses to deal with deviations of the live performer from the score. While [9] approaches score following using a Dynamic Bayesian Network and particle filtering for inference, better modelling rests and tempo changes than other approaches; [10] proposes Needleman-Wunsch time warping (NWTW), a pure dynamic programming method to align music recordings that contain structural differences. [11 and [12] focus on better alignment in the scenario where multiple performances are available, showing some promising results.

There remains a number of ***research challenges and opportunities***. Traditional approaches to music alignment typically rely on hand-crafted features, which often fail to generalise to different instruments, acoustic environments and recording conditions. We aim to address this feature engineering bottleneck by employing deep neural networks, which can capture both low-level features of relevance to the alignment task and higher-level mappings between

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

feature sequences of corresponding performances. We plan to explore data augmentation and semi-supervised learning methods to aid training, since deep networks are typically data hungry. As an example of knowledge-driven development, one challenge faced by [8] is complex piano music played with a lot of expressive freedom in terms of tempo changes. Hence, two ways are proposed in [13] to estimate the current tempo of a performance on-line, and how to use this information to improve the alignment. They incorporate tempo models which illustrate different possible performance strategies. This helps making the tracking algorithm more robust to on-the-fly structural changes.

As part of QMUL3, we plan to exploit the knowledge available to us through understanding of the way recording techniques have changed over the decades, by using techniques such as random filters for data augmentation for training the alignment models. Musical domain knowledge such as the fact that pianos are generally not perfectly in tune motivates further aspects of our data augmentation strategy.

5.4 QMUL4 "Robust Timbre Analysis for Query by Vocal Imitation"

The ***goal*** of this project is to study how a high-resolution, careful analysis of sound timbre can help sound designers and musicians to effortlessly find a certain desired sound by imitating it vocally. We merge traditional timbre analysis and deep learning algorithms to link vocal imitations to the sound being emulated.

Regarding the ***state of the art***, sound timbre has been studied for many decades, with perception studies and signal processing techniques making useful models out of it. Recently, a book was published on traditional timbre analysis [1] detailing all these strategies, which this project follows closely. However, timbre being such a complex psychoacoustical phenomenon, the success of traditional signal processing algorithms has been limited to some extent, as these heavily rely on human-engineered features and mostly lack transformation capabilities so as to achieve, for instance, timbre transfer between different musical instruments.

It has been discovered recently that generative models and deep learning methods in general are able to tackle these problems and come up with more robust and complete timbre models when enough data is available [2-5]. Furthermore, some promising attempts on merging deep learning with traditional timbre analysis have been made recently [6-7]. Deep learning has also been used in sound query by vocal imitation [8], although no explicit timbral analysis has been carried out on this front, which we believe to be of key relevance.

All in all, this project aims to further develop the idea of "deep + traditional" timbre analysis and apply it to query by vocal imitation, which raises a number of ***research challenges and opportunities***. Timbre is usually defined as the psychoacoustical attribute that is different from pitch, loudness and perceived duration. While we already have very strong correlations for these last three attributes with physical parameters (fundamental frequency, sound energy and empirical duration respectively), correlations with timbre perception are weaker, often resulting in a mixture of different parameters [1]. This makes the study of timbre especially challenging and in need of more complex models like the ones derived from deep learning architectures. Exploring these could potentially shed light on the nature of timbre and enhance the performance of algorithms for query by vocal imitation.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

5.5 QMUL5 "Adversarial attacks to understand deep learning models for music"

Deep learning models are black boxes which means that we do not know exactly what is being learnt by a deep learning model and whether it is similar to domain knowledge we have of the task. Our ***goal*** is to try and bridge the gap between the black box approach of deep learning with the vast wealth of domain knowledge in music.

There is a small body of work, at the current ***state of the art***, that does interpretability of deep learning layers [1-3].

The main ***challenge*** is to simplify domain knowledge into a set of objective measures in order to compare it to the features of the deep learning model. We need to figure out how to design experiments and ask the right questions to see if people with domain knowledge in music can understand the features being learnt by deep learning models.

5.6 UPF1 "Facilitating Interactive Music Exploration"

The ***goal*** of this project is to improve the music exploration process by replacing typically used discretized tags with a continuous semantic space learned by deep-learning autotaggers and utilizing reinforcement learning and interactivity to optimize the process individually for each user.

Regarding the current ***state of the art*** on this particular aspect, typical categories of tags that are used for exploration are genres and moods. In the new proposed open autotagging dataset [1] there are 3 defined categories: genre, mood/theme, and instruments. Moods and genres have been used before for exploration [2, 3], but not in conjunction.

The biggest ***challenge*** in this research is to understand the semantics of the continuous space that is learned by deep-learning autotaggers. Explainability of deep-learning systems is a field of research on its own, thus there are limited resources that can be spent on it in the context of this research.

5.7 UPF2 "Methods for Supporting Electronic Music Production with Large-Scale Sound Databases"

The ***goal*** of the project is to characterise the harmonic and rhythmic content of audio loops for providing music makers with a way to navigate loop databases. For this, we can employ knowledge from harmonic compatibility, rhythmic similarity and characterisation to retrieve loops according to high-level semantic characteristics which the users can understand.

Regarding the relevant ***state of the art***, an audio characterisation task which has seen significant research in MIR is key estimation. This technique has been used as a first step for identifying compatible songs in a music library, so that DJs can mix them harmonically. In [1], key is defined as *a system of relationships between a series of pitches having a tonic, or central pitch class, as its most important element*. Key and harmonic information are essential for browsing EMP databases. Music makers need tonal information to mix and layer sound files according to their tonal content [2].

However, the key only defines a large-scale harmonic compatibility metric and recent work towards harmonic mixing has focused on characterising the small-scale harmonic compatibility. Harmonic mixing has been addressed through three methods: key affinity, chroma vector

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

similarity and sensory dissonance minimisation [3]. Key affinity, a common method in music applications, uses the key of two songs and uses their distance in the Camelot Wheel as a measure of dissimilarity. Chroma similarity approaches such as [4] take the cosine distance between the chroma vectors of two songs as a measure of dissonance. In [5], the authors use sensory dissonance models determine the pitch-shift that maximises the consonance of the sum of two audio clips.

Rhythm takes a very important place in EDM and has a very strong connection with loops. The first step towards analysing rhythmic compatibility in loops is to have a reliable tempo estimation for the loop alignment. The common framework in tempo estimation approaches is proposed summarised in [6], comprising three stages: extracting relevant features from the audio, calculating periodicities and extracting the dominant one. In [7], by exploiting an inherent property of loops, that their duration should be a multiple of the duration of a single beat for the BPM estimate, the authors are able to create a reliable confidence measure for tempo estimation for loops.

Audio based approaches for rhythmic similarity have focused on using low-level representations of the sound, such as onset detection functions, to: i) derive a beat spectrum and use the largest spectral peaks to compare rhythmic content, which is the methodology applied in [8]; ii) create a quantised representation of the rhythm pattern to propose a metric for pattern coincidence and for syncopation group coincidence [9].

The developments sketched above present a number of ***research opportunities*** for the current project. Techniques for rhythmic and harmonic characterisation have been developed in the literature and are ready for being practically used. We want to implement and evaluate these techniques for retrieving loops in large-scale audio databases. These will be implemented in a loop retrieval system which can then be evaluated by music makers.

5.8 UPF3 "Identifying and Understanding Versions of Songs with Computational Approaches"

The main ***goal*** of this project is to develop systems that automatically identify different versions of a given song using scalable computational methods. From a knowledge-driven perspective, we aim to use non-linear time series analysis and time series motif discovery methods to obtain a better understanding of the relations that link various versions of a particular song.

The group of methods we call "knowledge-driven" are designed specifically considering the importance of domain knowledge in solving the version identification task. These methods are deterministic and do not provide variations in their results due to external factors; thus, there is no "learning" process involved. They generally use a feature post-processing step and a similarity estimation step to estimate whether a pair of songs are versions of each other or not.

Feature post-processing steps are designed mainly for achieving key, tempo and structure invariance, and for representing the data obtained from a feature extraction step in a form that is suitable for similarity calculations. Common examples of feature post-processing steps are creating Beat-Synchronous Features [1, 2], Optimal Transposition Index [3, 4], State-Space Representation [5] and 2D Fourier Transform [6, 7].

After feature extraction and post-processing, the last step of many version identification systems is the estimation of similarity between two inputs. Systems developed in recent years

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

pay more attention to post-processing techniques to obtain a scalable similarity estimation with basic distance calculations while alignment methods such as the Smith-Waterman algorithm [8, 5] provide better precision scores but suffer from lower retrieval speed.

Integrating and exploiting domain-specific knowledge for such a task seems to be a ***challenge***, but promising. For the version identification task, knowledge-driven methods can be seen as providing a "more accurate but slow" estimation. We believe that a system that can be used on an industrial scale should follow data-driven approaches for this task. Nevertheless, in the process of developing a data-driven solution, one must not ignore the significance of incorporating domain knowledge into the solution. Although there are many version identification systems proposed in previous research, the amount of research that focuses on understanding their similarities regarding various musical dimensions was not very significant to date. With techniques like non-linear time series analysis and time series motif discovery algorithms, we aim to understand the relations among versions in a better way.

5.9 TPT1 "Behavioural music data analytics"

The ***goal*** of the project is to improve music recommendation systems by integrating users' contextual information in the recommendation process, in order to provide the right recommendations at the right time.

As music can be listened to in various situations [1], there has been many studies on the relationship between music preferences and users' context. For example, North et al. [2] studied the influence of 17 different listening situations on music preferences. The study showed that music preferences are not only dependent on the emotional response they evoke, but also on how they affect the quality of the listening situation. This suggests that music preferences are strongly associated with the listening environment. They categorized these 17 different situations into: activity, localized subdued behaviour, spirituality, and social constraints. Additionally, they studied the relationship between these situations and some acoustic attributes of the music, such as loudness and rhythm, and found that they are associated with the listening situations.

A similar study [3], also categorized different listening contexts into 3 categories: personal, leisure, and work. They further expanded each category into subcategories that are more specific to the situation. For example, personal is categorized into three subcategories: personal - being (e.g. sleeping or waking up), personal - maintenance (e.g. cooking or shopping), and personal - travelling (e.g. driving or walking). Similarly, Sloboda [4] studied the functions or purpose of listening to music for different users. He found that users listen to music for different purposes, such as activity, e.g. waking up or exercising, and mood enhancement, to put in better mood or for motivation.

While the previous studies focused on investigating the different situations where users listen to music, other studies focused on the effect of the user situation and surrounding environment on their music choices, i.e. how the listening context related to preferred music style. Gillhofer and Schedl [5] studied the effect of users' context on the genre and mood of selected tracks. They studied the use of contextual information to predict song, artist, genre, and mood of the tracks. They found that context information could help in predicting the artist and genre, but fell behind in predicting the mood. Schedl et al. [6] studied the similarity of music in similar

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

geospatial contexts. They relied on tweets tagged with geolocation from the "Million Musical Tweets Dataset" [7] to link music to their geolocation and to study their similarity. Other studies investigated the effect of more global contextual information, i.e. not frequently changing, such as cultural information [8-10] or user demographics [11].

The previous studies show no uniform approach to identifying which contextual information to consider when working with music. This leads to a sparse literature that is difficult to link together and restrains future research from building on top of previous studies. This opens up a lot of ***challenges and opportunities*** for the present PhD project.

By investigating the previous studies, we find that there is no common definition or taxonomy of relevant contextual information in music consumption. Hence, identifying the relevant contexts and building a taxonomy of contexts and how they relate to each other is one important challenge in this project. This would help in formalizing the problem in the research community and help future work in building on top of previous work.

Similarity, there are no available standard datasets for this problem, specifically, comprising tracks labelled with their context classes. This is another challenge that is important for future research to have a baseline and a dataset to compare new results with.

5.10 TPT2 "Voice Models for Lead Vocal Extraction and Lyrics Alignment"

Several of the project ***goals*** relate to knowledge-driven aspects of the project. The first one is to extract information from a lyrics transcript that is useful for singing voice separation. Specifically, this means to produce representations of an extra-musical information source such as text that can be exploited by audio source separation methods. This requires also to address text-to-audio alignment issues because a separation system needs to know which part of the lyrics is useful for a certain part of the processed audio signal. The second goal is to exploit this additional information in order to improve performance on the singing voice separation task. This requires to find means to include prior knowledge into state of the art data-driven singing voice separation methods.

Regarding the current ***state of the art,*** a recent and comprehensive overview of lead and accompaniment separation in music [2] organizes the different approaches in two main categories. The first category comprises model-based approaches. They exploit specific knowledge about the lead source, which often is the singing voice if present, about the accompaniment, or about both. The second category comprises data-driven approaches, which make use of machine learning to learn the separation task on large databases. Both kinds of methods come with their strengths and weaknesses. While model-based methods do not require much training data, they make strong assumptions about the source signals to be separated such as harmonicity, stationarity, repetitiveness, a certain pitch-curve, etc. Those assumptions lead to increased separation quality as long as they are valid. However, in cases where they are violated separation quality decreases. Data-driven approaches avoid making assumptions but they need a considerable amount of training data, which is often not easy to obtain. Moreover, those approaches often lack explainability and interpretability, which makes them more difficult to handle in a research context [2]. It should be mentioned that the two categories are not mutually exclusive.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

In the context of this document, model-based source separation methods, in particular those using specific side-information about the signal to separate, are most relevant and reviewed in this section. Model-based approaches have been the state-of-the-art for a long time until recently, when data-driven approaches have shown better performance [2]. However, additional signal information remains relevant as it has the potential to make data-driven methods more robust. Examples for such signal-based side-information, that can improve separation quality [3], are the score [4], the pitch [5], the text transcript [6, 7], or examples generated by users [8].

A musical score contains global information such as which instruments and notes are contained in a piece of music and local information such as when certain notes are played. A strong assumption is that aligned scores are available. Automatic alignment methods exist using chroma-features and alignment through dynamic time warping or Hidden Markov Models (HMMs) [4]. Difficulties lie in the fact that the score is open to interpretation by the performer and no absolute frequency information is contained. Score information has, for example, been used for initialization of Nonnegative Matrix Factorization (NMF) based models [9, 10]. Scores with only rough alignment require models to cope with uncertainty in the derived information. They have been used with Generalized Coupled Tensor Factorization (GCTF) [11] and deep learning based approaches [12]. In the latter case, the score is used to enforce structure on representations learned by the model. Convolutional Neural Networks (CNNs) are used in [13] for score-informed source separation of classical music together with a score following system to obtain coarsely aligned scores.

If the pitch is known, it can be used to inform the separation of the corresponding source as well. There are several main melody or pitch detection algorithms based on HMM [14], on NMF [15], or a combination of NMF and deep learning [16]. The pitch-information has mainly been exploited for the construction of harmonic masks. The pitch is assumed to be the fundamental frequency and assuming perfect harmonicity the harmonics can be derived [5, 17, 18, 19].

Also text has been investigated as side-information for source separation. For example, Nonnegative Matrix Partial Co-Factorization has been used for text-informed speech [6] and singing voice [7] separation. The voice in the mixture was modelled by an extended source-filter model [1]. A speech audio signal is generated from the text with speech synthesis. After aligning the mixture and the synthesized speech with Dynamic Time Warping (DTW), the model is optimized through multiplicative update rules while exploiting the fact that the same spectral envelopes manifest in the voice of the mixture and the synthesized voice in the same order. Performance is, however, limited by the alignment quality. In [20], a Deep Neural Network (DNN) is used for text-informed speech enhancement with noisy speech features and text-features are the input. That means, in contrast to the NMF based method above, the input features come from two different domains and the DNN learns a common representation. However, the text also needs to be aligned, which is done by a GMM-HMM speech recognition system.

The state of the art implies several new ***challenges and research opportunities.*** Prior knowledge about the source signal to be separated has successfully been used to improve source separation. As outlined above, one major challenge of including prior knowledge into source separation systems is to time-align the additional information with the audio signal. Furthermore, the additional information often comes from a different modality than audio,

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

which puts additional demands on the separation system, especially when it comes to learning representations. Until now, the alignment, the design of representations, and the actual separation were performed independently in most cases. Deep learning based approaches might offer the opportunity to perform these three steps jointly leading to more efficient usage of the extra knowledge. While today's state of the art data-driven methods achieve very good performance without extra information, they depend on the diversity of training data in order to generalize well. Including prior knowledge has therefore still the potential to make state of the art approaches more robust.

5.11 TPT3 "Multimodal movie music track remastering"

The ***goal*** of the project is to perform multi-modal/multi-view music source separation/enhancement which exploits previously not considered modalities such as the user's attention to the instrument to separate. In particular, we want to characterize the user's attention in terms of their brain response to a musical stimulus.

The relevant ***state of the art*** in the field, regarding the knowledge-driven aspect, is that informed source separation exploits all the available prior information about the sources and the mixing process along with the audio signal [1] and was proven to enhance the source separation process especially for music. Many works have been proposed. For instance, it has been proven that music score [2] and speech text [3] lead to a better separation. Visual features, such as the motion of sound sources [4] and lip motion analysis [5] are other examples of information that can be used imitating the human capability to exploit both audio and visual features to enhance speech recognition in a cocktail party scenario. Only a few works have been proposed in the last years that combines source separation/enhancement with auditory attention decoding characterized by EEG recordings [6-8]. However, they all focus on attention to speech stimuli: the core idea is to separate each sound source and use them to identify and enhance the attended speaker.

This offers a unique new ***research opportunity*****:** we aim to take advantage of these recent works in order to develop a new form of informed music source separation approach which can be referred as *neuro-steered music source separation*. This task, to my knowledge, was never addressed before. This case is particularly interesting because the knowledge provided to the source separation algorithms is extra-musical (it is not directly related to the music data) and is directly given by the user interacting with the music (it is a physiological response).

5.12 TPT4: "Context-driven Music Transformation"

The ***goal*** of the project is to enable transforming music in terms of artistic style. Specifically, the goal is to modify the style of a piece while preserving some of its original content. The target style can be pre-defined, taken from an example (a piece in the target style) or based on some other variables or constraints (e.g. to adapt the piece to a particular user, a movie scene or a gameplay situation).

This project is concerned mostly with generic data-driven methods which should not require prior knowledge about the target style. To the extent that knowledge-based techniques will play a role in the project, the most relevant pieces of work that define the ***state of the art*** here

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

are works which constrain the problem using prior knowledge such as on (re-)harmonization [1-2], and expressive performance generation [3-5].

The main ***challenge*** with respect to knowledge-driven approaches is that they often do not generalize to new domains/styles (for which this knowledge might be unavailable). On the other hand, fully data-driven approaches are limited by the fact that the training data is not aligned, possibly making the problem ill-posed. To address these issues, we propose a different way to inject knowledge into the system, which is to generate synthetic parallel training data [6]. This data enables supervised or semi-supervised learning, overcoming the problem described above. In this case, the knowledge comes from the creators of the software we use to generate the data.

5.13 TPT5 "Conditional Generation of Audio Using Neural Networks and its Application to Music Production"

From a knowledge-driven perspective, the ***goal*** of project TPT5 is to study the impact of different time-frequency representations in audio generation using Neural Networks. We will explore different audio representations as these may improve the learning process and reduce the complexity and depth of the networks.

Audio signals carry overwhelming amounts of data in which relevant information for a specific task is often challenging to find. In general, feeding in sparse representations of the audio content, with few coefficients revealing the information we are looking for (i.e., representations where most coefficients are zero), yields faster training, and reduces the complexity of the architectures. ***State-of-the-art*** works on audio generation have made use of different ways of representing the audio content. Some have focused on raw audio data [1, 2] or mel-scaled STFT [3], others have attempted the generation of parameters for a synthesizer [4]. However, there are still many audio representations to be explored. Currently, we are studying the use of an invertible CQT [5], and we consider for the future other sparser audio representations such as wavelet-based scattering transforms [6].

There are a number of open ***research challenges***. After reviewing the leading works in the literature, we believe that most of them have either focused on raw audio, STFT, or mel-scaled spectrograms. Also, in many works using these time-frequency representations, the phase information is often neglected in favor of iterative reconstruction algorithms such as Griffin-Lim. As mentioned above, other approaches have used the neural network to parametrize a synthesizer. In its most raw representation, audio turns out to be very costly to process. At the other extreme, sparsity may yield a loss of information. For these reasons, we believe there are still many ways of representing the audio content to be explored. The main challenges are the complexity and computational effort behind many of these representation techniques.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

5.14 JKU1 "Large-scale Multi-modal Music Search and Retrieval without Symbolic Representations"

The ***goal*** of project JKU1 is to develop and improve methods for the automatic structuring and cross-linking of large multi-modal music collections. A special challenge is to work directly from sheet music images and audio recordings, without being able to rely on symbolic representations (machine-readable formats), which would be an unrealistic assumption in many real-world application scenarios.

The ***state of the art*** in multimodal score/audio retrieval consists mainly in three methodologies, which are summarized as follows. The first and traditional approach is to convert both visual and acoustic domains (sheet music images and audio recordings, respectively) into a common representation based on chroma features, also known as the chromagram. Balke et al. [3-4] use this methodology to propose a fully automated processing pipeline that matches sheet music queries to corresponding items within a database of audio recordings. In this methodology, the audio recordings in the collection are first transformed into a database of chromagrams. Following that, the sheet music query is also converted into the same chroma representation using an optical music recognition (OMR) software. Then, a matching procedure based on dynamic time warping (DTW) is applied to compute matching curves and finally to create a ranked list that represents the similarity between the query and each item in the dataset. Despite its novelty and promising results, this approach may be quite costly when dealing with large collections of music, therefore it calls for more efficient indexing methods.

The second main approach is based on symbolic fingerprinting, where both visual and acoustic representations are converted into a symbolic music domain and therefore transformed into compact and discriminative features. Arzt et al. [1-2] propose a method to address the score identification task, when given an audio snippet query, followed by retrieving the exact position in the score corresponding to the audio query. Moreover, the proposed approach is also tempo- and transposition-invariant, in order to address the problem of dealing with different versions of the same piece. This is done by defining the fingerprint as, taking three note events, an encoded triple calculated over the pitch and time instant relations of these note events. The symbolic fingerprinting method is described as follows. First, the score database is converted into an indexed fingerprint database from MIDI files. Following that, in the acoustic domain, the audio query is converted into a group of symbolic fingerprints via an automatic music transcription algorithm based on recurrent neural networks [6]. Finally, the audio query fingerprints are matched to each instance within the score fingerprint database and the best match is taken as the corresponding score. As advantages of this approach, fingerprints are compact and significantly discriminative; however it strongly depends on accurate transcriptions from audio to symbolic domains, this being a complex task. In the visual domain, it also depends on the performance of OMR systems in case of dealing with sheet music directly from images.

Exploiting the recent advances in artificial intelligence for representation learning, the third and last approach is to learn correspondences between sheet music images and audio recordings directly from a multi-modal training set by means of deep neural networks [7-8]. This cross-modal network learns a joint embedding space from both modalities by minimizing the distance between corresponding sheet music snippet and audio excerpt pairs. After training the network, the query procedure is straightforward and consists of mapping short segments of

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

both the query and the database into the embedding space. Then, the similarities between them are computed via their cosine distances, and the candidate from the database that is closer to the query is retrieved as the result. One limitation of this approach is that, since the learning stage is supervised, good generalization to new and unseen data requires a large amount of training data, comprising different instrumentation and musical styles. On the other hand, it does not require the definition of handcrafted representations, since it is based on an end-to-end approach. Also, the query can be in either domain, i.e. it is possible to retrieve scores given an audio query and vice-versa, which is a limitation of the afore-reviewed methods that only work in one direction.

From the current state of the art as described above, and from our goal of extending this to a massive scale, follow a number of specific ***challenges and opportunities*** for original research. In general, there will be no symbolic, machine-readable scores (e.g., MIDI or MusicXML) available for the overwhelming majority of the pieces. Therefore the main challenging aspect of our research is to provide solutions which rely directly on score images and audio recordings. As we plan to use as a starting point the approach introduced by Dorfer et al. [7-8], we identify an important limitation which is the *fixed field of view* on both audio and sheet music snippets. While not being a problem on the visual side, global and local tempo changes are commonly found in audio recordings, as a consequence of musical liberties taken and expressive choices made by the performers. Therefore one important aspect of our research is to make sure the algorithms to be developed are tempo-invariant, in order to not lose performance when dealing with different *tempi*.

Another characteristic of the embedding space learning approach is that the retrieval procedure is snippet-wise, i.e. the query is segmented and the corresponding excerpts are independently retrieved, based on a ranked histogram algorithm. We believe that taking into account knowledge about inherent *temporal dependencies* in music could result in a more robust and fast retrieval. Therefore we highlight this as a potential immediate topic to be studied.

Lastly, in the context of calculating fine-grained alignments of multiple performances to sheet music in big multi-modal music collections, which may be useful for tasks such as score-based listening and comparison or musicological analysis, we believe that general knowledge of typical *musical section structure* will become important. For instance, it could improve bar, staff and note head retrieval from sheet images, a task crucial for alignment applications. Generally, knowledge about different musical styles and notation or performance conventions will be crucial to permit our algorithms to correctly interpret various patterns they learn from the different modalities, and to relate these to each other in meaningful ways.

5.15 JKU2 "Live Tracking and Synchronisation of Complex Musical Works via Multi-modal Analysis"

The ***goal*** of project JKU2 is to develop a live music tracking system which follows in real time an opera performance along its respective sheet music. The tracker will be based on a multi-modal analysis relying on audio and video recordings in order to ensure robustness and accuracy.

The ***state of the art*** in live tracking of music has improved considerably in previous years. At the heart of this is the problem of audio-to-score alignment, which is the task of synchronizing an audio recording of a musical piece with the corresponding symbolic score; if done in real time, this is known as score following. The first approach presented in the 1980s [1] synchronized

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

MIDI representations of a monophonic melody with the corresponding score. Since then, synchronization evolved from a relying on symbolic representations of music [2,3] to raw audio [4,5]. While at first the focus was on monophonic music [6,7], currently available methods can follow highly polyphonic and complex orchestral music [8,9].

Synchronizing two different entities requires there to be a common space between them. The sheet music represents a set of notes, each described by an onset and an offset time. Although a recent approach considers score following directly on the graphical sheet music [10] (see also Project JKU1), for this project we will rely on symbolic representations. In our case, we will use MIDI files, generated from MusicXML or MEI files.

Common acoustic features can be extracted from both audio and scores. A usual representation, called semigram [11], groups spectral energies coming from the Fast Fourier Transform into semi-tones, adding a best accuracy in low frequencies. Another representation, called chromagram [12], groups harmonic pitches into 12 chromatic classes, adding robustness to timbre differences. In addition to the pitch and when the melody synchronization fails, some systems focus on tempo [13,14].

Two different approaches for score following are commonly used in the current state of the art: probabilistic methods and dynamic time warping based approaches. Among probabilistic models, Hidden Markov Models (HMM) [7,15] are widely used in the literature. For a real-time application, models must have few variables. They are composed of a hidden variable, the played chord, and an observable variable, the acoustic feature. In some cases, a tempo variable [16] is added to the system and ghosts states [4,17] can be used to make the system robust to mismatches. The alignment path can be extracted in real time with an efficient forward algorithm. Some recent techniques use particle filtering [13,16] to reduce the complexity. Other approaches use conditional random fields [18,19] which have no hypothesis concerning the observations, but only model conditional probabilities of the hidden variables given the observation sequence.

The second approach is based on the Dynamic Time Warping (DTW) algorithm. DTW is an efficient method to find the optimal alignment between two sequences. The path can be found inside a cumulative score matrix reflecting local scores between sequences. Because of the quadratic time and space complexity, some adaptations have been proposed to make this method usable in real-time applications [20-23]. Tempo information can also be added into the tracking [24].

In order to get a robust and accurate live music tracker, we also propose to detect relevant events not only in the audio but also in the video. Detecting events in audio is a way to recognize and identify acoustic scenes. In an opera context, voice, applause or even instrument detection can inform the system about playing parts, recitatives and act endings. Detecting such events in audio is a common machine learning task, and we will try to make use of existing approaches, where possible. For example, the AudioSet dataset [25] includes detection of music, male and female singing voice and applause.

In opera, the voice is widely present during the performance. Detecting voice, the gender of the voice or the lyrics can be relevant to robustly read the sheet music. As the voice has highly specific frequency trajectories, they are visually detectable in a spectrogram. The most common features are the Mel-Frequency Cepstral Coefficients (MFCC), used e.g. by Lehner [26] for an

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

on-line algorithm. More recent work uses fluctograms combined with spectral contraction and spectral flatness features [27]. These mid-level features have already been applied to classical opera recordings [28].

The gender of the voice can also be an interesting clue to detect. Male and female voices are different in terms of fundamental frequency F0, bandwidth and amplitude. Perceptual Linear Prediction (PLP) coefficients can be efficiently used for this task [29]. Tracking spoken language can be also relevant. The singing voice of an opera singer in a mixed signal is hard to transcribe, but it might be useful for recitative parts.

Applause may happen at specific times during a performance, for example at the entrance of the conductor, at ends of acts and at the end of the performance. Additionally, applause might happen after a singer finishes a well-known aria. Thus its detection gives strong indications about specific events. From a frequency point of view, applause has very irregular patterns in the spectrogram. A combination of MFCCs and other low level features such as spectral centroid (representing the gravity center of the spectrum), spectral spread (representing the variance), spectral flux (representing the dissimilarity of succeeding frames) and spectral flatness (a measure representing the deviation of a spectrum from a flat shape) can be efficiently used for applause detection [30].

Finally the challenging task to follow an opera instead of an orchestra performance involves the use of all modalities available during live streaming. The video can contain information which is difficult to find in audio. Indeed the curtain call, the set changes, the number of actors on stage or the playing/non-playing activity of musicians can more easily be detected in analyzing images. Face and gender recognition systems can be useful to cluster actors on stage. The main features in the field are eigenfaces and fisherfaces [29,31].

The task of tracking a complex acoustic event such as an opera presents us with specific ***challenges and opportunities*** for original research. In particular, the scenario of tracking entire live operas is more complex than anything tackled so far in this field. For instance, we are now faced with a mixture of voice, music, and possibly other sounds which is often interrupted by intermissions. Music tracking systems such as those discussed above are not designed to be robust to these problems which will seriously complicate the tracking task. In this project, we will approach the problem via new multi-modal tracking algorithms which will be based on audio and video input, on specialized features and detectors for important, recurring events, and on hierarchical tracking schemes, which try to make sense of this data and relate it to the score description.

As partner, the Vienna State Opera (VSO) will support this project. Over the last few years, they have invested heavily in technology for live streaming to make their performances accessible to larger audiences. The VSO will provide this project with the data needed for our research. This includes (multi-channel) audio recordings of performances, video from multiple angles, subtitles and scores.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

6. Conclusion

Starting from a general notion of extra-musical knowledge and its possible roles in music information research, this document has given an overview of knowledge-related aspects as they arise in the sub-projects making up MIP-Frontiers. As can be seen, there is indeed a myriad of different ways in which general musical and extra-musical knowledge can be of help in solving complex music processing tasks. Beyond documenting the focus and progress of the project to the outside world and the project reviewers, the present report will also turn out to be a valuable intra-project resource, demonstrating to the ESRs the importance of knowledge-driven approaches to MIR research, and the multitude of ways in which knowledge, if available, can be exploited to improve data-driven, machine-learning-based solutions.

7. References

7.1 QMUL1 "Representation Learning in Singing Voice"

[1] Wilkins, Julia, et al. "VocalSet: A Singing Voice Dataset." ISMIR 2018.

[2] Smule Vocal Performances (multiple songs) Dataset, "https://ccrma.stanford.edu/damp/," accessed July 2018.

[3] Kruspe, Anna M. "Bootstrapping a System for Phoneme Recognition and Keyword Spotting in Unaccompanied Singing." ISMIR 2016.

[4] Gupta, Chitralekha, et al. "Semi-supervised Lyrics and Solo-singing Alignment." ISMIR. 2018.

[5] Dabike, G. R., "Barker, J. Automatic Lyric Transcription from Karaoke Vocal Tracks: Resources and a Baseline System" Interspeech 2019.

[6] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

7.2 QMUL2 "Improving Polyphonic Transcription through Instrument Recognition and Source Separation"

[1] S. Essid, G. Richard and B. David, "Musical instrument recognition on solo performances," in European Signal Processing Conference, 2004.

[2] A. Diment, P. Rajan, T. Heittola and T. Virtanen, "Modified Group Delay Feature for Musical Instrument Recognition," in Proceedings of the International Symposium on Computer Music Multidisciplinary Research, Marseille, 2013.

[3] S. Makino, Audio Source Separation, 1 ed., Springer International Publishing, 2018.

[4] Z. Rafii and B. Pardo, "REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation," IEEE Transactions on Audio, Speech and Languague Processing, vol. 21, pp. 73-84, 1 2013.

[5] Z. Rafii, A. Duan and B. Pardo, "Combining Rhythm-Based and Pitch-Based Methods for Background and Melody Separation," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 22, pp. 1884-1893, 12 2014.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

[6] S. e. a. Sofianos, "H-Semantics: A Hybrid Approach to Singing Voice Separation," Journal of the Audio Engineering Society, vol. 60, pp. 831-841, 10 2012.

[7] C. Lordelo, E. Benetos, S. Dixon and S. Ahlbäck, "Investigating kernel shapes and skip connections for deep learning-based harmonic-percussive separation," arXiv e-prints, p. arXiv:1905.01899, 5 2019.

7.3 QMUL3 "Leveraging User Interaction to Learn Performance Tracking Music Alignment"

[1] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing, 26(1):43{49, 1978.

[2] Ning Hu, Roger B Dannenberg, and George Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on., pages 185-188. IEEE, 2003.

[3] Meinard Müller, Frank Kurth, and Tido Röder. Towards an efficient algorithm for automatic score-to-audio synchronization. In ISMIR, 2004.

[4] Simon Dixon. An on-line time warping algorithm for tracking musical performances. In IJCAI, pages 1727{1728, 2005.

[5] Meinard Müller, Henning Mattes, and Frank Kurth. An efficient multiscale approach to audio synchronization. In ISMIR, pages 192{197. Citeseer, 2006.

[6] Andreas Arzt, Gerhard Widmer, and Simon Dixon. Automatic page turning for musicians via real-time machine listening. In ECAI, pages 241{245, 2008.

[7] Meinard Müller, Verena Konz, Andi Scharfstein, Sebastian Ewert, and Michael Clausen. Towards automated extraction of tempo parameters from expressive music recordings. In ISMIR, pages 69{74, 2009.

[8] Andreas Arzt and Gerhard Widmer. Towards effective 'any-time'music tracking. In Proceedings of the 2010 conference on STAIRS 2010: Proceedings of the Fifth Starting AI Researchers' Symposium, pages 24{36. IOS Press, 2010.

[9] Filip Korzeniowski, Florian Krebs, Andreas Arzt, and Gerhard Widmer. Tracking rests and tempo changes: Improved score following with particle filters. In ICMC, 2013.

[10] Maarten Grachten, Martin Gasser, Andreas Arzt, and Gerhard Widmer. Automatic alignment of music performances with structural differences. In Proceedings of 14th International Society for Music Information Retrieval Conference (ISMIR. Citeseer, 2013.

[11] Andreas Arzt and Gerhard Widmer. Real-time music tracking using multiple performances as a reference. In ISMIR, pages 357{363. Citeseer, 2015.

[12] Siying Wang, Sebastian Ewert, and Simon Dixon. Robust and efficient joint alignment of multiple musical performances. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(11):2132{2145, 2016.

[13] Andreas Arzt and Gerhard Widmer. Simple tempo models for real-time music tracking. In Proceedings of the Sound and Music Computing Conference (SMC), 2010.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

7.4 QMUL4 "Robust Timbre Analysis for Query by Vocal Imitation"

[1] Siedenburg, K., Saitis, C., McAdams, S., Popper, A.N., Fay. "Timbre: Acoustics, Perception, and Cognition." *Springer* (May 2019).

[2] Huang, Sicong, et al. "Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer." *arXiv preprint arXiv:1811.09620* (2018).

[3] Engel, Jesse, et al. "Gansynth: Adversarial neural audio synthesis." *arXiv preprint arXiv:1902.08710* (2019).

[4] Engel, Jesse, et al. "Neural audio synthesis of musical notes with wavenet autoencoders." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

[5] Luo, Yin-Jyun, Kat Agres, and Dorien Herremans. "Learning Disentangled Representations of Timbre and Pitch for Musical Instrument Sounds Using Gaussian Mixture Variational Autoencoders." *arXiv preprint arXiv:1906.08152* (2019).

[6] Esling, Philippe, Axel Chemla-Romeu-Santos, and Adrien Bitton. "Generative timbre spaces with variational audio synthesis." *Proceedings of the International Conference on Digital Audio Effects (DAFx)*. 2018.

[7] Esling, Philippe, Axel Chemla-Romeu-Santos, and Adrien Bitton. "Bridging Audio Analysis, Perception and Synthesis with Perceptually-regularized Variational Timbre Spaces." *ISMIR*. 2018.

[8] Zhang, Yichi, Bryan Pardo, and Zhiyao Duan. "Siamese Style Convolutional Neural Networks for Sound Search by Vocal Imitation." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.2 (2018): 429-441.

7.5 QMUL5 "Adversarial attacks to understand deep learning models for music"

[1] Choi, K., Kim, J., Fazekas, G., & Sandler, M. (2015). *Auralisation of Deep Convolutional Neural Networks: Listening to Learned Features.* International Society of Music Information Retrieval (ISMIR), Late-Breaking Demo.

[2] Mishra, S., Sturm, B., & Dixon, S. (2017). *Local Interpretable Model-Agnostic Explanations For Music Content Analysis.* International Society for Music Information Retrieval (ISMIR).

[3] Mishra, S., Sturm, B., & Dixon, S. (2018). *Understanding a Deep Machine Listening Model Through Feature Inversion.* International Society of Music Information Retrieval (ISMIR).

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

7.6 UPF1 "Facilitating Interactive Music Exploration"

[1] Bogdanov, D., Won, M., Tovstogan, P., Porter, A. & Serra, X. The MTG-Jamendo dataset for automatic music tagging. In Proceedings of the Machine Learning for Music Discovery Workshop, 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (2019). URL http://mtg.upf.edu/node/3957.

[2] Andjelkovic, I., Parra, D. & O'Donovan, J. Moodplay: Interactive mood-based music discovery and recommendation. In Vassileva, J., Blustein, J., Aroyo, L. & D'Mello, S. K. (eds.) Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016, 275{279 (ACM, 2016). URL https://doi.org/10.1145/2930238.2930280.

[3] Pampalk, E. Islands of music: Analysis, organization, and visualization of music archives. Master's thesis, Vienna University of Technology (2001). URL http://www.ofai.at/~elias.pampalk/music/pampalk_summary.pdf.

7.7 UPF2 "Methods for Supporting Electronic Music Production with Large-Scale Sound Databases"

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

[2] Angel Faraldo, Emilia Gómez, Sergi Jordà, and Perfecto Herrera. "Key estimation in electronic dance music". In 38th European Conference on Information Retrieval, pages 335–347, Padua, Italy, 2016. Springer-Verlag.

[3] Gilberto Bernardes, Diogo Cocharro, Marcelo Caetano, Carlos Guedes, and Matthew E.P. Davies. "A multi-level tonal interval space for modelling pitch relatedness and musical consonance". Journal of New Music Research, 45(4):281–294, 2016

[4] M. E. P. Davies, P. Hamel, K. Yoshii, and M. Goto. "Automashupper: Automatic creation of multi-song music mashups." IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(12):1726–1737, Dec 2014.

[5] Roman B. Gebhardt, Matthew E. P. Davies, and Bernhard U. Seeber. "Psychoacoustic approaches for harmonic music mixing". Applied Sciences, 6(5), 2016.

[6] Sebastian Bock, Florian Krebs, and Gerhard Widmer. "Accurate tempo estimation based on recurrent neural networks and resonating comb filters." In Proceedings of the 16th International Society for Music Information Retrieval Conference, 2015.

[7] Font F, Serra X. Tempo estimation for music loops and a simple confidence measure. In Proceedings of the 17th International Society for Music Information Retrieval Conference; 2016

[8] Gerard Roma and Xavier Serra. "Music performance by discovering community loops". In WAC - 1st Web Audio Conference, Paris, 2015.

[9] Daniel Gómez-Marín, Sergi Jordà, and Perfecto Herrera. "Pad and sad: Two awareness-weighted rhythmic similarity distances". In Proceedings of the 16th International Society for Music Information Retrieval Conference, 2015.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

7.8 UPF3 "Identifying and Understanding Versions of Songs with Computational Approaches"

[1] Ellis, D. P. W. & Poliner, G. E. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In Proc. of 2007 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), IV–1429–IV–1432 (Honolulu, HI, USA, 2007).

[2] Bertin-Mahieux, T. & Ellis, D. P. W. Large-scale cover song recognition using hashed chroma landmarks. In 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 117–120 (New Paltz, NY, USA, 2011).

[3] Serrà, J., Gómez, E., Herrera, P. & Serra, X. Chroma binary similarity and local alignment applied to cover song identification. IEEE Transactions on Audio, Speech, and Language Processing 16, 1138–1151 (2008).

[4] Silva, D. F., Yeh, C.-C. M., Batista, G. E. A. P. A. & Keogh, E. J. SiMPle: Assessing music similarity using subsequences joins. In Proc. of 17th Int. Conf. on Music Information Retrieval (ISMIR), 23–29 (New York City, NY, USA, 2016).

[5] Serrà, J., Serra, X. & Andrzejak, R. G. Cross recurrence quantification for cover song identification. New Journal of Physics 11, 093017 (2009).

[6] Bertin-Mahieux, T. & Ellis, D. P. W. Large-scale cover song recognition using the 2D Fourier Transform magnitude. In Proc. of 13th Int. Conf. on Music Information Retrieval (ISMIR), 241–246 (Porto, Portugal, 2012).

[7] Humphrey, E., Nieto, O. & Bello, J. Data driven and discriminative projections for large-scale cover song identification. In Proc. of 14th Int. Conf. on Music Information Retrieval (ISMIR), 4–9 (Curitiba, Brazil, 2013).

[8] Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. Journal of Molecular Biology 147, 195–197 (1981).

7.9 TPT1 "Behavioural music data analytics"

[1] Alinka E Greasley and Alexandra Lamont. Exploring engagement with music in everyday life using experience sampling methodology. *Musicae Scientiae*, 15(1):45–71, 2011.

[2] Adrian C North and David J Hargreaves. Situational influences on reported musical preference. *Psychomusicology*: *A Journal of Research in Music Cognition*, 15(1-2):30, 1996.

[3] John A Sloboda, Susan A O'Neill, and Antonia Ivaldi. Functions of music in everyday life: An exploratory study using the experience sampling method. *Musicae scientiae*, 5(1):9–32, 2001.

[4] John A Sloboda. Everyday uses of music listening: A preliminary study. Music, mind and science, pages 354–369, 1999.

[5] Michael Gillhofer and Markus Schedl. Iron Maiden while jogging, Debussy for dinner? In *International Conference on Multimedia Modeling*, pages380–391. Springer, 2015.

[6] Markus Schedl, Andreu Vall, and Katayoun Farrahi. User geospatial context for music recommendation in microblogs. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 987–990. ACM, 2014.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

[7] David Hauger, Markus Schedl, Andrej Košir, and Marko Tkalčič. The million musical tweet dataset: What we can learn from microblogs. *International Society for Music Information Retrieval*, 2013.

[8] Martin Pichl, Eva Zangerle, Günther Specht, and Markus Schedl. Min-ing culture-specific music listening behavior from social media data. In *2017 IEEE International Symposium on Multimedia (ISM)*, pages 208–215.IEEE, 2017.

[9] Marcin Skowron, Florian Lemmerich, Bruce Ferwerda, and Markus Schedl. Predicting genre preferences from cultural and socio-economic factors for music retrieval. *In European Conference on Information Retrieval*, pages 561–567. Springer, 2017.

[10] Bruce Ferwerda and Markus Schedl. Investigating the relationship between diversity in music consumption behavior and cultural dimensions: A cross-country analysis. *In UMAP (Extended Proceedings)*, 2016.

[11] Thomas Krismayer, Markus Schedl, Peter Knees, and Rick Rabiser. Predicting user demographics from music listening information. *Multimedia Tools and Applications*, 78(3):2897–2920, 2019.

7.10 TPT2 "Voice Models for Lead Vocal Extraction and Lyrics Alignment"

[1] Fant, Gunnar. *Acoustic theory of speech production*. No. 2. Walter de Gruyter, 1970.

[2] Z. Rafii, A. Liutkus, F.-R. Stöter, S. Ioannis Mimilakis, D. Fitzgerald, and B. Pardo. An overview of lead and accompaniment separation in music. *IEEE Transactions on Audio, Speech, and Language Processing,* 26(8):1307, 2018.

[3] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot. From blind to guided audio source separation: How models and side information can improve the separation of sound. *IEEE Signal Processing Magazine,* 31(3):107–115, 2014.

[4] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley. Score-informed source separation for musical audio recordings: An overview. *IEEE Signal Processing Magazine,* 31(3):116–124, 2014.

[5] T. Virtanen, A. Mesaros, and M. Ryynänen. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In *SAPA@INTERSPEECH*, pages 17–22, 2008.

[6] L. Le Magoarou, A. Ozerov, and N. Q. Duong. Text-informed audio source separation. example-based approach using non-negative matrix partial cofactorization. *Journal of Signal Processing Systems,* 79(2):117–131, 2015.

[7] L. Le Magoarou, A. Ozerov, and Q. K. N. Duong. Method of singing voice separation from an audio mixture and corresponding apparatus, Dec. 31 2015. *US Patent App.* 14/748,164.

[8] D. Fitzgerald. User assisted separation using tensor factorisations. *European Signal Processing Conference (EUSIPCO),* 2412–2416, 2012.

[9] J. Fritsch and M. D. Plumbley. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. *IEEE International Conference on Acoustics, Speech and Signal Processing,* pages 888–891, 2013.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

[10] A. Liutkus, J. L. Durrieu, L. Daudet, and G. Richard. An overview of informed audio source separation. In *International Workshop on Image Analysis for Multimedia Interactive Services,* 2013.

[11] U. Simsekli and A. T. Cemgil. Score guided musical source separation using Generalized Coupled Tensor Factorization. *European Signal Processing Conference (EUSIPCO),* 2639–2643, 2012.

[12] S. Ewert and M. B. Sandler. Structured dropout for weak label and multiinstance learning and its application to score-informed source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing,* pages 2277–2281, 2017.

[13] M. Miron, J. Janer Mestres, and E. Gómez Gutiérrez. Monaural score-informed source separation for classical music using convolutional neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference,* 2017.

[14] C.-L. Hsu, L.-Y. Chen, J.-S. R. Jang, and H.-J. Li. Singing pitch extraction from monaural polyphonic songs by contextual audio modeling and singing harmonic enhancement. In *Proceedings of the International Society for Music Information Retrieval Conference,* pages 201–206. Citeseer, 2009.

[15] J.-L. Durrieu, G. Richard, and B. David. An iterative approach to monaural musical mixture de-soloing. In *IEEE International Conference on Acoustics, Speech and Signal Processing,* pages 105–108. IEEE, 2009.

[16] D. Basaran, S. Essid, and G. Peeters. Main melody extraction with source-filter nmf and crnn. In *Proceedings of the International Society for Music Information Retrieval Conference,* 2018.

[17] Y. Ikemiya, K. Itoyama, and K. Yoshii. Singing voice separation and vocal f0 estimation based on mutual combination of robust principal component analysis and subharmonic summation. *IEEE Transactions on Audio, Speech, and Language Processing,* 24(11):2084–2095, 2016.

[18] Y. Ikemiya, K. Yoshii, and K. Itoyama. Singing voice analysis and editing based on mutually dependent f0 estimation and source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing,* pages 574–578. IEEE, 2015.

[19] Z. Rafii, Z. Duan, and B. Pardo. Combining rhythm-based and pitch-based methods for background and melody separation. *IEEE Transactions on Audio, Speech, and Language Processing,* 22(12):1884–1893, 2014.

[20] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani. Text-informed speech enhancement with deep neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association,* 2015.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

7.11 TPT3 "Multimodal movie music track remastering"

[1] A. Liutkus, J. Durrieu, L. Daudet, and G. Richard. "An overview of informed audio source separation". In 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pages 1–4. IEEE, 2013.

[2] S. Ewert, B. Pardo, M. Müller, and M. D Plumbley. "Score-informed source separation for musical audio recordings: An overview." IEEE Signal Processing Magazine, 31(3):116–124, 2014.

[3] L. Le Magoarou, A. Ozerov, and N. QK Duong. "Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization". Journal of Signal Processing Systems, 79(2):117–131, 2015.

[4] S. Parekh, S. Essid, A. Ozerov, N. QK Duong, P. Pérez, and G. Richard. "Guiding audio source separation by video object information". In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pages 61–65. IEEE, 2017.

[5] T. Afouras, J. S. Chung, and A. Zisserman. "The conversation: Deep audio-visual speech enhancement". arXiv preprint arXiv:1804.04121, 2018.

[6] S. Van Eyndhoven, T. Francart, and A.r Bertrand. "Eeg-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses". IEEE Trans. Biomed. Engineering, 64(5):1045–1056, 2017.

[7] N. Das, S. Van Eyndhoven, T. Francart, and A. Bertrand. "Eeg-based attention-driven speech enhancement for noisy speech mixtures using n-fold multi-channel wiener filters". In 25th European Signal Processing Conference (EUSIPCO), pages 1660–1664. IEEE, 2017.

[8] J. A O'sullivan, A. J Power, N. Mesgarani, S. Rajaram, J. J Foxe, B. G Shinn-Cunningham, M. Slaney, S. A Shamma, and E. C Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cerebral Cortex, 25(7):1697–1706, 2014.

7.12 TPT4 "Context-driven Music Transformation"

[1] Pachet, François and Pierre Roy. "Non-Conformant Harmonization: the Real Book in the Style of Take 6." ICCC (2014).

[2] Hadjeres, Gaëtan, Jason Sakellariou and François Pachet. "Style Imitation and Chord Invention in Polyphonic Music with Exponential Families." ArXiv abs/1609.05152 (2016).

[3] Widmer, Gerhard, Sebastian Flossmann and Maarten Grachten. "YQX Plays Chopin." AI Magazine 30 (2009): 35-48.

[4] Flossmann, Sebastian and Gerhard Widmer. "Toward a multilevel model of expressive piano performance." (2011).

[5] Malik, Iman and Carl Henrik Ek. "Neural Translation of Musical Style." ArXiv abs/1708.03535 (2017).

[6] Cífka, Ondrej, Umut Simsekli and Gaël Richard. "Supervised Symbolic Music Style Translation Using Synthetic Data." ArXiv abs/1907.02265 (2019).

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

7.13 TPT5 "Conditional Generation of Audio Using Neural Networks and its Application to Music Production"

[1] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals,Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. "Wavenet: A generative model for raw audio." In The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016, page 125, 2016.

[2] Sungwon Kim, Sang-gil Lee, Jongyoon Song, and Sungroh Yoon. "Flowavenet : A generative flow for raw audio." CoRR, abs/1811.02155, 2018

[3] Sean Vasquez and Mike Lewis. "Melnet: A generative model for audio in the frequency domain." CoRR, abs/1906.01083, 2019.

[4] Merlijn Blaauw and Jordi Bonada. "A neural parametric singing synthesizer." 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017, pages 4001–4005 ,2017

[5] Gino Angelo Velasco, Nicki Holighaus, Monika Doerfler, and Thomas Grill. "Constructing an invertible constant-q transform with nonstationary gabor frames." Proceedings of the 14th International Conference on Digital Audio Effects, DAFx2011, 09 2011.

[6] Vincent Lostanlen and Stephane Mallat. "Wavelet scattering on the pitch spiral." CoRR, abs/1601.00287, 2016.

7.14 JKU1 "Large-scale Multi-modal Music Search and Retrieval without Symbolic Representations"

[1] A. Arzt, S. Böck, and G. Widmer, "Fast identification of piece and score position via symbolic fingerprinting", in Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR), Porto, Portugal, 2012.

[2] A. Arzt, G. Widmer, and R. Sonnleitner, "Tempo- and transposition-invariant identification of piece and score position", in Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR), Taipei, Taiwan, 2014.

[3] S. Balke, S. P. Achankunju, and M. Müller, "Matching musical themes based on noisy OCR and OMR input", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 2015.

[4] S. Balke, V. Arifi-Müller, L. Lamprecht, and M. Müller, "Retrieving audio recordings using musical themes", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016.

[5] H. Barlow and S. Morgenstern, A Dictionary of Musical Themes. Crown Publishers, Inc., 3rd ed., 1975.

[6] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks", in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012.

[7] M. Dorfer, A. Arzt, and G. Widmer, "Learning audio-sheet music correspondences for score identification and offline alignment", in Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 2017.

[8] M. Dorfer, J. j. Hajič, A. Arzt, H. Frostel, and G. Widmer, "Learning audio-sheet music correspondences for cross-modal retrieval and piece identification", Transactions of the International Society for Music Information Retrieval, vol. 1, no. 1, pp. 22-33, 2018.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

7.15 JKU2 ("Live Tracking and Synchronisation of Complex Musical Works via Multi-modal Analysis")

[1] Dannenberg, R. B. (1984, October). An on-line algorithm for real-time accompaniment. In *ICMC* (Vol. 84, pp. 193-198).

[2] Vercoe, B. (1984). The synthetic performer in the context of live performance. In *Proceedings of International Computer Music Conference* (pp. 199-200).

[3] Jordanous, A., & Smaill, A. (2009). Investigating the role of score following in automatic musical accompaniment. *Journal of New Music Research*, *38*(2), 197-209.

[4] Orio, N., & Déchelle, F. (2001). Score following using spectral analysis and hidden Markov models.

[5] Müller, M., Mattes, H., & Kurth, F. (2006, October). An efficient multiscale approach to audio synchronization. In *ISMIR* (Vol. 546, pp. 192-197).

[6] Raphael, C. (2010, June). Music Plus One and Machine Learning. In *ICML* (pp. 21-28).

[7] Cano, P., Loscos, A., & Bonada, J. (1999, October). Score-Performance Matching Using HMMs. In *ICMC*.

[8] Arzt, A., & Widmer, G. (2015, October). Real-Time Music Tracking Using Multiple Performances as a Reference. In *ISMIR* (pp. 357-363).

[9] Prockup, M., Grunberg, D., Hrybyk, A., & Kim, Y. E. (2013). Orchestral performance companion: Using real-time audio to score alignment. *IEEE MultiMedia*, *20*(2), 52-60.

[10] Dorfer, M., Henkel, F., & Widmer, G. (2018). Learning to listen, read, and follow: Score following as a reinforcement learning game. *arXiv preprint arXiv:1807.06391*.

[11] İzmirli, Ö., & Dannenberg, R. B. (2010, August). Understanding Features and Distance Functions for Music Sequence Alignment. In *ISMIR* (pp. 411-416).

[12] Hu, N., Dannenberg, R. B., & Tzanetakis, G. (2003, October). Polyphonic audio matching and alignment for music retrieval. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)* (pp. 185-188). IEEE.

[13] Otsuka, T., Nakadai, K., Takahashi, T., Ogata, T., & Okuno, H. (2011). Real-time audio-to-score alignment using particle filter for coplayer music robots. *EURASIP Journal on Advances in Signal Processing*, *2011*(1), 384651.

[14] Grosche, P., Müller, M., & Kurth, F. (2010, March). Cyclic tempogram—A mid-level tempo representation for musicsignals. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5522-5525). IEEE.

[15] 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.

[16] Korzeniowski, F., Krebs, F., Arzt, A., & Widmer, G. (2013). Tracking rests and tempo changes: Improved score following with particle filters. In *ICMC*.

[17] Montecchio, N., & Orio, N. (2009). A Discrete Filter Bank Approach to Audio to Score Matching for Polyphonic Music. In *ISMIR* (pp. 495-500).

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

[18] Joder, C., Essid, S., & Richard, G. (2011). A conditional random field framework for robust and scalable audio-to-score matching. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(8), 2385-2397.

[19] Sako, S., Yamamoto, R., & Kitamura, T. (2014, August). Ryry: A real-time score-following automatic accompaniment playback system capable of real performances with errors, repeats and jumps. In *International Conference on Active Media Technology* (pp. 134-145). Springer, Cham.

[20] Dixon, S. (2005, September). Live tracking of musical performances using on-line time warping. In *Proceedings of the 8th International Conference on Digital Audio Effects* (pp. 92-97).

[21] Macrae, R., & Dixon, S. (2010, August). Accurate Real-time Windowed Time Warping. In *ISMIR* (pp. 423-428).

[22] Arzt, A., Widmer, G., & Dixon, S. (2008, July). Automatic Page Turning for Musicians via Real-Time Machine Listening. In *ECAI* (pp. 241-245).

[23] Fremerey, C., Müller, M., & Clausen, M. (2010). Handling Repeats and Jumps in Score-performance Synchronization. In *ISMIR* (pp. 243-248).

[24] Arzt, A., & Widmer, G. (2010, July). Simple tempo models for real-time music tracking. In *Proceedings of the Sound and Music Computing Conference (SMC)*.

[25] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... & Ritter, M. (2017, March). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 776-780). IEEE.

[26] Lehner, B., Sonnleitner, R., & Widmer, G. (2013, November). Towards Light-Weight, Real-Time-Capable Singing Voice Detection. In *ISMIR* (pp. 53-58).

[27] Lehner, B., Schluter, J., & Widmer, G. (2018). Online, loudness-invariant vocal detection in mixed music signals. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *26*(8), 1369-1380.

[28] Dittmar, C., Lehner, B., Prätzlich, T., Müller, M., & Widmer, G. (2015, October). Cross-Version Singing Voice Detection in Classical Opera Recordings. In *ISMIR* (pp. 618-624).

[29] Pronobis, M. (2008). *Integrating audio and vision for robust automatic gender recognition* (No. REP_WORK). Idiap.

[30] Uhle, C. (2011). Applause sound detection. *Journal of the Audio Engineering Society*, *59*(4), 213-224.

[31] Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7), 711-720.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.