New Frontiers in Music Information Processing (MIP-Frontiers)

Grant Agreement Number: 765068

- Title: State of the art, challenges and potential of data-driven approaches in MIR
- Lead Beneficiary: UPF (Universitat Pompeu Fabra)
- 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 data-driven aspects – the focus of the present report – the main message is that MIR research needs to face the ever-growing amount of data available to be processed and made sense of; and address music problems for which new audio signal processing and machine learning methodologies will have to be developed.

The preparation of this report started with an internal discussion, within the consortium, on what data-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 will need to use large datasets and deep-learning methodologies, but they target different aspects of data-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 on how the quality of datasets, data augmentation strategies, identifying suitable architectures and the interpretability of developed systems are important for MIR and for their specific problems. This report has already been, and will be, a valuable resource, demonstrating to the ESRs and to the MIR community the importance of data-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 data-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 "data-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. The first and most relevant one is the development of data-driven methods and system-level solutions of interest to a wide variety of companies working within the audio and music field. For this goal we need large and appropriate corpora/datasets, state of the art audio/music feature analysis tools, and machine learning and evaluation strategies adequate for the specific problems identified.

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 data-driven approaches to MIR problems. This report will focus on the data-driven research topics as they emerge in the specific research projects (PhD theses) tackled in MIP-Frontiers. To this end, the next section (Section 4) gives an overview of which of these projects are particularly related to data-driven methods, 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. Data-driven methods in MIP-Frontiers

4.1 The concept of data-driven in MIR

Music data is in most cases complex, multimodal, and it exists in very large quantities (e.g. in the scale of hundreds of thousands of items for music pieces in diverse modalities, or tens of millions in the case of audio files or tags). In addition, music is increasingly available in data streams rather than data sets, and the characterization of music data can evolve with time (e.g. tag annotations are constantly evolving, sometimes even in an adverse way). These Big Data characteristics (very large amounts, streaming, non-stationarity) imply a number of challenges for MIR, such as data acquisition, dealing with weakly structured data formats, scalability, online (and real-time) learning, semi-supervised learning, iterative learning and model updates, learning from sparse data, learning with only positive examples and learning with uncertainty.

The rest of this section lists those projects in MIP-Frontiers where data-related aspects as defined above are particularly important, and briefly explains why and in what way. For each of these projects, Section 5 will then describe the current state of the art and corresponding challenges and opportunities.

[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.

4.2 MIP-Frontiers projects with data-driven aspects

QMUL1: "Representation Learning in Singing Voice"

What this project aims at from a data-driven perspective is to create training datasets from weakly or not labelled singing voice data. This project takes advantage of the data that DoReMir provides, which consists of over one million recordings of real-world singing. The research adapts unsupervised & semi-supervised learning methods to learn new latent representations 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.

However, only some indications of characteristics of sounds that affect our perception of timbre are known. We still do not fully understand how much each physical characteristic correlates with the other and how much they actually contribute to timbre perception. Thus, finding an explicit mathematical formula of how we exactly perceive timbre using the physical properties of instruments and of sound waves is a highly complex and unsolved task. Therefore, the project is going to exploit timbre representations learned directly from data by using machine/deep learning techniques.

QMUL3: "Leveraging user interaction to learn performance tracking"

The QMUL3 project focuses on music alignment. Music alignment aims at providing a way to navigate among multiple representations of music in a unified manner, lending itself applicable to a myriad of domains like music education, performance, enhanced listening, automatic accompaniment and so on. The project has a heavy data-driven component, since we aim to develop deep learning based approaches to music alignment, which require large datasets containing labelled alignments.

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

The type of data associated with this project is given in pairs of audio files, comprising the original target sounds and their vocal imitations, given by different users. Publicly available datasets in this field are scarce and often composed by a small number of users.

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

Project QMUL5 generates adversarial attacks on deep learning models for music and tries to analyze the features in the deep learning model that are exploited to generate these adversarial examples. We want to establish a link between the type of data used to train a model and the robustness of the model.

[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.

UPF1: "Facilitating Interactive Music Exploration"

Project UPF1 aims to utilize audio and tag annotations to facilitate the process of music exploration. The data will be used to train a deep-learning autotagging system to learn embeddings which will be used as anchors for exploration in the continuous semantic 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, data-driven techniques will be used to identify the instruments which are present in each loop.

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. We aim to incorporate state-of-the-art data-driven techniques from the machine learning community into this line of research in order to develop systems that can be used on an industrial scale.

TPT1: "Behavioral music data analytics"

Project TPT1 aims at improving music recommendation systems by integrating the user's contextual information in the recommendation process. The user's contextual information is defined as the external factors that affect their music preferences in any given time. For example, user activity, location, or time of the day are considered as contextual information that changes users' 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, we need to consider the audio content of music tracks to provide the right recommendation at the right time. While most current recommendation systems rely on collaborative filtering approaches, using audio content is an alternative approach that is being studied frequently nowadays. Hence, our data-driven approach considers integrating the audio content information along with context and user information to provide better recommendations.

TPT2: "Voice models for lead vocal extraction and lyrics alignment"

Audio source separation is the task of extracting individual sound sources such as lead vocals from a mixture. Project TPT2 aims at developing robust audio source separation methods for singing voice extraction. Due to the high complexity of musical audio mixtures, where sound sources are usually highly correlated in time and frequency and may be modified by (non-linear) audio effects, machine learning techniques are required for such a separation task. This means, the task is learned on training data. Consequently, methods for singing voice separation can be considered as data-driven.

TPT3: "Multimodal movie music track remastering"

The specific research goal of project TPT3 is to enhance source separation algorithms taking advantage of the data extracted from the user while he/she is listening to polyphonic music. Specifically, this data is represented by the brain responses of the user to musical stimuli collected using electroencephalographic 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.

TPT4: "Context-driven music transformation"

The aim of project TPT4 is to enable transforming music in terms of artistic style. While simple modifications can be designed by hand, complex transformations such as changing the musical genre will most likely be best performed by machine learning models trained on large music datasets.

TPT5: "Conditional generation of audio using neural networks and its application to music mroduction"

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 crosslinking of large multi-modal music collections, with a focus on audio recordings and sheet music images (acoustic and visual domains, respectively). In addition to the use of additional higher-level knowledge of music which could be used to improve alignment, identification and retrieval, solving this will require massive amounts of musical data. These comprise audio recordings and score images in various representations.

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 make use of data provided by the Vienna State Opera, partner of this project, and coming from different modalities (such as audio and video recordings from opera performances) in order to detect events which are useful to achieve robustness.

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

5.1 QMUL1 "Representation Learning in Singing Voice"

Through unsupervised learning techniques and taking advantage of the size and variety in the dataset provided by DoReMir, the specific task is to learn latent representations of the singing voice and further create clean and refined subsets from big unlabelled datasets for the purpose of training supervised learning algorithms.

State of the Art. Many modern deep learning studies apply deep unsupervised learning to learn features and latent representations from audio data. The advantage of learning representations is that the plentiful unlabelled data can be utilized to obtain new representations of the data that are potentially better than hand-crafted features [1]. The authors of [2] propose to use Sequence-to-sequence Autoencoder (SA) and its extension for unsupervised learning of Audio Word2Vec to obtain vector representations for audio segments. They show SA can learn 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.

representations describing the sequential structures of the audio segments. The authors of the paper [3] proposes a combination of Semantic Variational Autoencoders with RNNs (SVAE - RNN) to obtain global latent representations of audio data. In the construction of the unsupervised representation learning system, this concept of autoencoders will be exploited.

Challenges and Opportunities. The main challenge in unsupervised learning is the size of data required. This project will take advantage of the DoReMir dataset for this purpose. The recordings in the dataset are collected from users from over 100 countries, using a mobile music transcription application which gives itself a great potential to represent real-world singing data. The DAMP singing datasets (available for research, released by Smule) [4,5,6,7] also provide big potential for the specific task.

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

The project explores the inter-dependencies between instrument recognition and source separation tasks with the final goal of improving polyphonic instrument transcription. This can be achieved by proposing the creation of a system capable of automatically learning timbre-related features for identifying the different sounds that are being played and capable of separating the music signal into multiple sources based on the learned timbres.

State of the Art. Even though instrument recognition and source separation are two tightly connected tasks, in the specialized literature they are usually addressed separately. The former is usually seen as a multi-label classification task, which is performed and evaluated either at a frame-level, note-level or clip-level, while the latter is defined as an unmix operation, where one tries to separate the audio signal into multiple instrumental sources.

With the recent release of new mid and large scale datasets such as MusicNET [1], which contains 330 freely-licensed classical music recordings written for 11 instruments, along with over 1 million annotated labels indicating the precise time of each note in every recording and the instrument that plays them, fully data-driven supervised deep-learning techniques for instrument recognition started to arise in the literature. For instance, [2] allies the constant-Q transform with pitch information from a multi-pitch estimation algorithm to perform frame-level instrument recognition using a convolutional network. Later, the authors improved the recognition method by using a multi-task learning approach, trying to jointly predict the instrument class and the pitch of the notes. Other important examples are [3] and [4]. Both works perform instrument recognition in polyphonic music in clip-level basis. Their main difference is that [3] uses the mel-frequency spectrogram of the clip as input to the network while [4] uses an end-to-end approach, i.e., the time-domain audio clip is used as input for the system to directly pinpoint the multiple instruments in the mixture.

Regarding source separation, with the increasingly utilisation of data-driven approaches and the follow up improvements of the state-of-the-art performance in closely related music information retrieval tasks such as multi-pitch estimation and instrument tracking, deep learning methods also started to make break-throughs in the area of audio source separation [5-8]. However, a limitation of most methods is the fact that they are designed to tackle predefined separation sub-tasks, i.e., they rely on extracting only a specific instrument from each other or are focused on performing the separation only in particular music genres. Some

[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.

examples are methods for melody and vocal extraction [9] and for lead and accompaniment separation [10]. Moreover, those types of methods do not usually generalise well to different types of music or to other instrumental sources.

More recently, the data-driven state-of-the-art methods started to use deep neural networks capable of learning long-term time relationships of the audio recording. Some examples are the utilisation of Recurrent Neural Networks (RNN) and Long-Short Term Memory (LSTM) networks, which have been shown to improve the separation [11,12].

Challenges and Opportunities. Despite the constantly increasing availability of instrument recognition and source separation datasets, they are usually designed for just one of those tasks. So, they lack the necessary annotations for performing the other task. Therefore, we need to create our own dataset to tackle these two tasks together and then apply them to improving automatic transcription.

Moreover, novel deep-learning architectures and user-driven methods will also be developed as result of the project. One example is a proposal of a novel architecture for performing harmonic-percussive source separation [13] that was already accepted for publication.

5.3 QMUL3 "Leveraging user interaction to learn performance tracking"

The overall goal of the project QMUL3 is to develop robust alignment methods which have significant coverage and can adapt to the setting they are being employed in. This involves training models on different kinds of audio data, to improve the coverage, and additional fine-tuning of models based on local context and by leveraging user data. There is a significant requirement of alignment data in these terms to build strong alignment models. The data driven aspects of this project overlap significantly with the user driven aspects, since we will be leveraging a combination of labelled data (pre-annotated) and contextual data (from the users' local context).

State of the Art. While most of the traditional approaches to alignment are based on Dynamic Time Warping [1], machine learning approaches have started showing promising results to music alignment. [2] is one of the earliest methods to apply deep learning for alignment. They employ an end-to-end multi-modal convolutional neural network trained on sheet music images and audio spectrograms of the corresponding snippets. [3] approaches the alignment task via automatic music transcription using Recurrent Neural Networks. [4] and [5] are recent works focusing on end-to-end alignment learning. While they work on snippets of performance aligned to an image, we would like to work on direct alignment between the audio and the symbolic domain computed in an end-to-end fashion.

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 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.

[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.

There have been a few approaches which employ deep learning for the task of music synchronization. The major ones among these approaches include [2] and [4]. Although these works are initial explorations in the direction of using DNN's for alignment, they prove that deep learning is a viable method to be chosen for the alignment task. While all of these works focus on generating alignments for a snippet of audio to an image of the sheet music using multi-modal training, we propose to employ these methods in a situation where the scores of the different pieces are available. This is since the Tido platform focuses on the use case where users already know which piece they are going to play, and they have a score file available already for that. This will enable us to employ the information present in the score and override the potential drop in accuracy introduced by the (lack of) robustness of Optical Music Recognition (in cases where the image is not available directly). A challenge to be faced is the availability of training data comprising manually labelled alignments. We will leverage existing datasets and potentially augment them with data obtained from the Tido platform. Another challenge to be addressed is to optimally deal with the multimodal nature of the datasets and the long-term dependencies present in the inputs. We will attempt to tackle these challenges using novel architectures which are capable of handling long-term dependencies as well as multi-modal input configurations.

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.

State of the Art. We will be querying two types of data through vocal imitation: sound samples (recordings) and synthesizer parameter configurations. The first part of the project, comprising the two first years of research, focuses on the former, specifically on the retrieval of musical, environmental, and synth preset sounds (audio recordings, not parameters).

We have started analysing drums (drum sample query by vocal imitation) and, while some datasets are already available [1-3], they mostly focus on pure sound classification instead of practical querying. That brought us to record our own dataset of vocal percussion [4] in order to fit the querying paradigm. This was done by first telling the algorithm how we vocally imitate, for instance, a kick drum (recording around twenty isolated utterances) and then using that data to identify kick drum imitations in vocal percussion improvisations, which include imitations of several drums. The utterances in these improvisation files are usually performed with different intonations, accentuations and durations, while preserving most of their timbral information.

Regarding other musical and environmental sounds, the VocalSketch dataset [5] comprises thousands of crowdsourced imitations of these. This dataset also features vocal imitations of several synth sounds and includes their respective parameter configurations.

C**hallenges and Opportunities.** The dataset we recorded [4] will potentially bring robustness to the algorithms for drum sample query by vocal percussion. The VocalSketch dataset [5] seems ideal to study when it comes to the rest of the query sounds and synth parameter configurations, and therefore there is no need to record a new dataset for those.

[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"

We compare the features learned by deep learning models that use standard datasets and use data augmentation. In MIR data augmentation is a popular technique to counter the issue of limited datasets. However, data augmentation has mixed results. We hope to shed some light on how data augmentation changes the features learnt by a deep learning model to provide more clarity on what data augmentation does.

State of the Art. Data augmentation techniques were shown to be effective in singing voice detection. We are using the singing voice model created by [4].

Challenges and Opportunities. The main challenge of our work is to identify the best form of data augmentation for a task and provide some explanations as to why that particular data augmentation technique is suitable for the task.

5.6 UPF1 "Facilitating Interactive Music Exploration"

The goal of this project is to improve the music exploration process by replacing typically used discrete 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.

State of the Art. Recommendation systems normally utilize either collaborative filtering [1] or a content-based approach [2], or mixture of both. State of the art deep-learning autotagger systems [3] commonly use mel-spectrograms as input, several layers of convolutional neural networks (CNNs) and dense layers (one or more) before the output [4]. There are several open datasets for autotagging [5, 6, 7], however, due to quality and standardization issues, we have proposed a new MTG-Jamendo dataset [8].

Challenges and Opportunities. Nowadays there are multiple autotagging datasets, however, it is difficult to share audio that is copyrighted. Thus, we have proposed a new dataset in collaboration with Jamendo that contains audio that is licensed by creative commons. With the audio openly available and the annotation quality curated by Jamendo, it presents new opportunities for researchers in autotagging and deep-learning.

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

The data-driven goal of this project is to automatically identify the instruments present in an audio loop, so as to allow searching by instrument in large-scale sound databases, by incorporating data-driven approaches with musically motivated design choices.

State of the Art. Automatic instrument classification is a classic task in Music Information Retrieval. However, research in this field has been mostly performed on clean and unprocessed sounds in small datasets [1,2]. On the other hand, the sounds provided by databases with audio for music production may also contain "production-ready" sounds, with audio effects applied on them, and sounds recorded in very different recording conditions.

[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.

Automatic classification of instruments has been targeted in a knowledge-driven manner, where handcrafted features developed by experts were used together with a classification algorithm. New approaches employ data-learned features by using Convolutional Neural Networks for this classification, which is the case of [3,4,5]. In [3] the author uses musically motivated filters in the CNN architecture, reducing the number of trainable parameters by a significant amount, while maintaining the classification accuracy comparable to other state-of-the-art models.

In order to increase the generalisation of a model further than the data provided to it, one possible approach is to use data augmentation. This approach can be described as applying deformations to a collection of training samples, in a way that the correct labels can still be deduced. These approaches were used for classifying sounds in [6,7].

Challenges and Opportunities. Some commercial and community databases of loops for electronic music production offer audio loops which are annotated with the predominant instrument in the loop. By collecting a large enough dataset from these databases, we can use deep learning approaches for automatically identifying the predominant instrument. Techniques for data augmentation using digital audio effects will be employed to improve the robustness and the classification accuracy of a deep learning model for instrument classification.

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 data-driven perspective, we plan to explore both the advantages of deep learning approaches for this task and the trade-off between identification accuracy and retrieval speed in order to develop a highly scalable system.

State of the art. In terms of data-driven methods for the version identification task, it is safe to say that the attention data-driven point of view getting is increasing as in many research communities due to many successful applications in recent years. We would like to note that the data-driven methods do not neglect domain knowledge, and the models they use are designed combining both domain knowledge and the successful techniques proposed by the machine learning community.

For data-driven methods, we mainly consider the deep learning approaches for the version identification task, and we look at two main aspects, namely types of architectures and types of approaches. In terms of architecture types, convolutional architectures [1, 2, 3, 4] consist of convolutional, pooling and linear layers while the recurrent architectures [5] consist of Recurrent Neural Network (RNN) variants such as Long Short-Term Memory (LSTM) as well as pooling and linear layers. Each having their advantages over the other, a best practice in terms of model architectures has not been agreed upon.

Another aspect for looking at the data-driven methods for this task would be the type of training approaches, namely classification and similarity. Examples of the classification approaches can be binary classification where the network is deciding whether two input songs are versions of each other or not [1], or multi-label classification where each version group is considered as a separate class [2]. The similarity approaches [3, 4, 5], on the other hand, aim to encode each

[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.

song into a latent representation using Siamese Networks, and the distance between two latent vectors would be the estimation of their similarity in terms of musical characteristics they share. Regarding previous works for this task, the similarity approaches appear to be more popular than the classification ones.

Challenges and opportunities. Based on the success of the data-driven methods, we aim to explore deep learning approaches for the version identification task in a detailed manner. Our goal is to introduce and integrate state-of-the-art techniques from the machine learning community into this line of research. By doing so, we aim to develop scalable systems that can be used in an industrial scale. During the project, we will investigate the current limitations of version identification systems from an industrial point of view, and by taking advantage of domain knowledge and the successful deep learning advancements, we expect to have an impact on both academia and industry. Moreover, the data we plan to curate during this project will have a significant impact for the community.

5.9 TPT1 "Behavioral music data analytics"

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

State of the Art. Several approaches have been proposed to improve the music recommendation process, and recommendation systems in general. They are often categorized depending on the approach and type of information used in the recommendation process [1,2]. One of the most common approaches is collaborative filtering, which relies on using the similarities between the users or items to find suitable recommendations. Another method of recommendation systems is content based recommendation (data-driven), which relies on using the actual content of the items, e.g. the audio content in the case of music, to recommend songs to users based on their listening history [3,4,5,6].

Other data-driven approaches rely on auto-tagging the tracks with relevant tags that could be used in searching, recommending, and filtering music according to the user's preferences. Recent approaches of auto-tagging rely on applying deep learning techniques on the raw audio data. Specifically, popular approaches apply convolutional neural networks (CNN) on the mel spectrograms of tracks [7,8]. However, the previous studies did not specifically study auto-tagging using contextual tags nor did they study the relationship between these contextual tags and audio content.

Challenges and Opportunities. Previous work has not studied the relationship between audio content and contextual information. One challenge is to identify to what degree the context of the user affects the choice of music style. Is there a dominant acoustic feature that prevails in each of these contexts? Are some of these contexts more influential in choosing specific music style than others? All of these questions require a joint study of audio content and user's context. This would be helpful in identifying the importance of certain contexts in the recommendation process and help in auto-tag tracks with their suitable context classes.

[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.

5.10 TPT2 "Voice models for lead vocal extraction and lyrics alignment"

The data related project goal is to explore training on multimodal data. In addition to the usual audio source separation training data consisting of audio mixtures and the corresponding individual source stems, this project also makes use of the corresponding lyrics transcripts. The latter also contains information about the singing voice but are from another modality (text vs. audio). The project researches possibilities to include the text data into the training procedure.

State of the art. A recent and comprehensive overview of lead and accompaniment separation in music [1] 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 [1]. It should be mentioned that the two categories are not mutually exclusive. In the context of this document, the data-driven approaches are relevant and will be reviewed in this section.

Data-driven separation systems usually learn a non-linear mapping between a mixture signal and spectral masks, which are used to filter the mixture in the Short Time Fourier Transform (STFT) domain to obtain the target source STFT. They can also learn a mapping to the target source directly, either in the STFT domain or the time domain.

The U-net [2] is a convolutional neural network (CNN) adopted from image segmentation to audio source separation. It operates on magnitude spectrograms and is able to recognize and use global as well as local characteristics thanks to its encoding-decoding architecture. On the encoding side, the spectrogram is processed by several strided 2D-convolution layers making the time and frequency scale coarser in deeper layers allowing to compute more global features. In the following decoding side, the feature maps go through several deconvolution layers, which upsample the feature maps back to the original dimensionality. The outcome is a feature map with the same size as the input spectrogram. The feature maps of the encoding side are concatenated to the corresponding feature maps of the decoder side with the same time-frequency resolution. This ensures that no information is lost in the encoding side and results in a nice gradient flow. The output is a soft mask that can be applied to the mixture spectrogram.

The multi-scale multi-band DenseNet [3] is another CNN based network for audio source separation using an encoder-decoder architecture. One difference to the U-Net is the use of dense blocks comprising several layers in which the output of all preceding layers is concatenated to the input of the next layer within one block. Another difference it that they apply different networks to different frequency bands to account for their different patterns and different importance. Another interesting deep learning approach is multi-task learning,

[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.

where a network is trained to perform two related tasks simultaneously. Stoller et al. [4] perform singing voice separation and singing voice detection jointly with such an approach. A U-Net like network [2] for singing voice separation is extended with a component that predicts vocal activity based on a hidden layer of the separation model. Thus, the relationship between annotated vocal activity and the voice source to be separated is modelled in an implicit way. This allows the network for example to be robust to temporally inaccurate labels. The resulting network can perform voice separation and detection at the same time. The loss function is a weighted sum of the loss functions of both tasks. The authors show that the performance on both tasks is improved compared to a single-task network with the same architecture.

Two end-to-end approaches for singing voice have been published recently. They take a mixture signal in the time domain as input and output the separate source signals in the time domain as well. The advantage is that the phase information is contained in the input features as opposed to magnitude spectrograms as input. Moreover, the difficult choice of STFT parameters is omitted this way. Stoller et al. [5] adapt the U-Net architecture to the one-dimensional case to operate directly on samples of a waveform input. The encoding-decoding structure, which here boils down to downsampling and upsampling layers, allows to derive local and global features just as in the 2-dimensional case explained above. This so-called Wave-U-Net achieves state-of-the-art performance for singing voice separation. The WaveNet [6], a generative model for raw audio, has been adapted for audio source separation as well [7]. The authors turn the network into a non-causal, discriminative model as has been done before for speech denoising [8]. That is, future samples are also used to predict the current sample. As opposed to the original WaveNet, which samples from a softmax output layer, the adapted version directly regresses the source waveforms. In terms of SDR and in listening tests the WaveNet-based approach performs slightly worse than the Wave-U-Net.

Challenges and Opportunities. The main data-related challenges of singing voice separation lie in the size and diversity of training data. There are some data sets that are widely used by the musical source separation community such as MUSDB18 [9]. To compare different methods, it is important to use the same data for training and testing across the research community. However, usually these data sets are rather small and do (of course) not contain examples from all possible music styles. In order to train models that generalize well to any style it is required to have diverse training data. As opposed to data sets for other MIR tasks, a source separation data set cannot easily be created by manual annotations as access to raw recording material with each instrument on a separate track is required. Those are not available for commercial music as it underlies copyright restrictions. Different researchers have access to different non-public additional training data, which they use to improve performance as it usually scales with the amount of training data. However, this hinders fair comparison of data-driven source separation approaches.

Another challenge regarding deep learning is the interpretability of results and models. It is not easy and sometimes impossible to understand how exactly the model makes its predictions or what exactly it has learned. Consequently, it is not straightforward to define the right actions to tackle shortcomings.

[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.

5.11 TPT3 "Multimodal movie music track remastering"

The goal of the project is to perform a multimodal/multiview 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.

State of the Art. Studying the problem at hand requires data of well-synchronized musical stimuli and corresponding brain responses which can only be acquired in a controlled sensory stimulation. There are only a few publicly available music-related EEG datasets acquired in such a way [1, 2, 3], but they were designed for a different purpose and the subjects were not asked to attend to any particular instrument. The only one where participants were asked to focus on an instrument while listening to polyphonic music, is the music BCI dataset used in [4]. However, it was specifically designed for studying ERP-based attention decoding. Our focus is instead on single-trial attention decoding techniques, targeting real music compositions. Consequently, we acquired our own dataset at Télécom ParisTech.

Challenges and Opportunities. Music data nowadays is available in very large quantities and the number and type of annotations are constantly increasing. However, this is not true when considering as annotation, the physiological response of the user who is listening to the music. Such annotations are expensive and time-consuming to obtain. This is the main reason why only a few and small datasets are available in this field.

We acquired our own dataset, but the amount of data collected is still not enough to develop reliable data-driven techniques to characterize the user's attention with respect to music.

5.12 TPT4 "Context-driven music transformation"

The aim 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).

State of the art. Style transformation can be approached as a 'domain translation' task, where the goal is to translate data between different domains. From a machine learning perspective, we can approach the task in a supervised or unsupervised manner, depending on the kind of training data. An example of a task where aligned data is readily available, and therefore a supervised approach is feasible, is machine translation (MT) between natural languages. On the other hand, music style translation is an example of a task where a sufficient amount of aligned data is difficult or impossible to collect.

With the recent advances in deep learning, the task of translation has become widely studied for other modalities such as images [1] [2], even enabling unsupervised image-to-image translation [3] [4]. Similar techniques were soon applied to text [5] and finally, with remarkable results, to music audio [6]. Attempts to employ these and other deep learning techniques to perform unsupervised translation of symbolic music [7] [8] [9] have been somewhat less successful.

[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.

It has also been proposed to perform audio style transfer using techniques developed for images [10]. Given the nature of the style representation, the work is concerned with the transfer of sound textures and timbre rather than more high-level features.

Challenges and opportunities. The style transformation techniques described above have, for the most part, not yet permitted to obtain results as compelling as those on images. Clearly, this poses a significant challenge, and more work is required to adapt them to the musical domain. It is also quite possible that purely data-driven approaches (relying purely on machine learning on existing music data) will not be sufficient, and that some knowledge will be needed in order to guide the machine learning systems to a reasonable solution. This is further discussed in the appropriate section of report D2.1 (knowledge-driven approaches).

5.13 TPT5 "Conditional generation of audio using neural networks and its application to music production"

From a data-driven perspective, the goal of the project is to perform audio synthesis using generative models by exploiting multi-track audio data and musical attribute information (e.g., instrument type, tempo), as conditioning information.

State of the Art. Generative models require an immense amount of training data to model the probability distribution of a certain process accurately. Some of the most remarkable works attempting audio generation with neural networks have made use of solo piano databases [1, 2], music extracted from video-games [3] or a synthetically generated single-note audio database [4], which has become a de facto standard for these types of tasks. Currently, we are using the latter database for preliminary experiments conditioning only on the music attribute information. However, for the particular task at sight, a considerable number of databases exist with multi-track or stem audios, as well as descriptive metadata, such as those enumerated in the Stage 0 report [5, 6, 7]. We will also consider a more extensive range of data obtained from Sony CSL's private collection, although for publishing purposes, we will focus on publicly available data.

Challenges and Opportunities. Exciting times are ahead of us given the constant increase of audio databases and the precision with which these are annotated. Many doors are opened for data-driven methods requiring large databases, such as the one under consideration. The main challenge that we have faced throughout this first stage of the project is computing power. The complexity and depth of the models, together with the amount of data needed for such complex tasks, requires enormous computing power (or a lot of time and patience). For this reason, and as explained in deliverable Stage 1, we may consider in the future, if required, the usage of cloud computing for training heavy models.

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

The goal of project JKU1 is to develop methods for the automatic structuring and cross-linking of large multi-modal music collections. Moreover, it is proposed to develop algorithms which do not rely on symbolic representations (machine-readable formats), since such collections are built directly from sheet music images and audio recordings.

[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.

State of the Art. The state of the art in multimodal score/audio retrieval consists mainly of 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, known as a 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.

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 (derived from MIDI data) 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.

Exploiting the recent advances in artificial intelligence for representation learning, the third and last approach learns 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. In these works, the MSMD dataset [8] is used for evaluation of results. This multi-modal dataset comprises a large number of precisely annotated solo piano pieces, for a total of more than a thousand pages of music and about 15 hours of aligned audio.

Challenges and Opportunities. From the current state of the art as described above, and from our goal of extending this to a massive scale, a number of specific challenges and opportunities for original research follow. Our motivation is to develop robust, scalable methods for supporting several related tasks: retrieval of one modality based on another (e.g. retrieval of audio recordings given score image queries); alignment of multiple performances to sheet music for purposes of score-based listening and comparison; and piece identification in unknown recordings, e.g., for automatic metadata provision. Since our goal is to extend state of the art methods to a massive scale, one crucial aspect of our research will be to identify, and possibly augment, potential data to be exploited. First, the MSMD dataset mentioned above is still a suitable starting point for our purposes. Although completely artificial, it could be re-rendered for different instrumentation and/or genres. Moreover, various creative and musically meaningful forms of *data augmentation* (which has proven to be an extremely effective method in many applications of deep learning) will have to be investigated. Of course, also ways of extending the number and diversity of musical pieces (and real interpretations of these) will be targeted. With respect to the use of real (real-world) score images, the IMSLP Petrucci Music Library (which contains over 400,000 scores and 50,000 recordings) is a promising online data source that will be investigated for this project.

[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.

5.15 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.

State of the Art. 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 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 to have 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), we will rely for this project on symbolic representations. In our case, we will use MIDI files, generated from MusicXML or MEI files.

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) [6,7] are widely used in the literature. For a real-time application, models must have few variables. They are composed by 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 be 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 useable in real-time applications [20-23]. Tempo information can also be added into the tracking [24].

Challenges and Opportunities. Our task is more difficult than anything attempted in this field before, which presents us with a number of new research challenges and opportunities, specifically related to the data aspect. 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. In this project, we will thus have to develop new multi-modal tracking algorithms using both audio and video input, and specialized features and detectors for important, recurring events, among others. Developing, optimizing, and evaluating these are very data-intensive tasks.

As our main source of real-world opera recordings, we will rely on the infrastructure that is already put in place at the Vienna State Opera (VSO) for the production of their regular live

[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.

streams. We will be able to use more than twenty microphones and eight cameras which record the performances. As a first test piece, we have selected the opera *Don Giovanni* by Mozart. For this piece, we have received, from the VSO, a set of development data, which consists of audio recordings with microphones placed at different locations, and video recordings of the whole piece with a single point of view of the stage, with multiple points of view with camera moves, and with the conductor view.

The analysis of the sheet music as a PDF is not considered in this project. (This is something that project JKU1 will likely work on; we hope to be able to benefit from that work in later phases of the project.) Instead, we will build on two digital representations to track the performance. The musicXML format includes onset and offset of each note; a piano roll representation can be extracted from that. The MEI (Music Encoding Initiative) format is a more complete representation, including context information and music annotations. Here, we are trying to obtain a complete encoding of the entire opera Don Giovanni from the Mozarteum Foundation in Salzburg. However, it may only be possible to obtain parts of the opera from this source. We will thus have to find ways of complementing this high-quality encoding with less complete representations of other parts of the opera (e.g., MIDI files from the Web, or annotated audio recordings).

The audio recordings are given as a multi-track recording and also as the final mastered version. They are in high quality and they are mainly focused on one type of instrument, suggesting an instrument recognition analysis. Finally, video recordings are given in a 4k definition. The main stage is recorded with a fixed point of view, one camera focuses on the conductor and the mastered video using different moving cameras is also given. Due to a high quality, all the data represents more than 700 GB, which constitutes a solid but also challenging data basis for our research.

An important concern in our project, however, is that we intend to follow an *open data policy.* Unfortunately (and understandably), the recording data provided by the Vienna State Opera cannot be made publicly available, due to rights issues. In order to still be able to make it possible for the research community to reproduce and build on our work, we will have to identify an alternative source of real-world opera recordings that can be openly shared. To this end, we plan to cooperate with the JKU Orchestra and the Anton Bruckner University of Music in Linz, around a semi-concertante performance of Mozart's opera *Die Zauberflöte*, which we will be permitted to record and distribute, along with any metadata and annotations that will be collected and produced in the process of our research. This adds an extra layer of effort to our work but will be important for scientific reasons.

[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

Data driven methods are at the core of all the projects to be carried out by all the ESRs.

As described in this document, the projects cover a wide variety of MIR problems, such as music recommendation, source separation, music transformation, sound synthesis, music transcription, music alignment, audio-based query, music exploration and cover song identification. The problems are not completely defined yet, but the ESRs have already been able to identify the relevant state of the art and challenges to be addressed.

All projects will require the use of large datasets and some projects will require ESRs to develop their own datasets. Issues of the quality of the datasets and data augmentation strategies will be required in some of projects.

Deep learning methodologies will be used in all the projects. Most of them will work on identifying architectures that are useful for music and for their specific problems. Issues of interpretability will also be an important concern for most projects.

7. References

7.1 QMUL1 "Representation Learning in Singing Voice"

References

[1] Langkvist, M., Karlsson, L. & Loutfi, A. A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Letters 42, 11{24 (2014)

[2] Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y. & Lee, L.-S. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. arXiv preprint arXiv:1603.00982 (2016)

[3] Jang, M., Seo, S. & Kang, P. Recurrent neural network-based semantic variational autoencoder for sequence-to-sequence learning. arXiv preprint arXiv:1802.03238 (2018)

[4] Kruspe, Anna M., and I. D. M. T. Fraunhofer. "Bootstrapping a System for Phoneme Recognition and Keyword Spotting in Unaccompanied Singing." ISMIR. 2016.

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

[6] Smule Sing! 300x30x2 Dataset, "https://ccrma.stanford.edu/damp/," accessed September 2018.

[7] Wager, Sanna, et al. "Intonation: A dataset of quality vocal performances refined by spectral clustering on pitch congruence." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 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.

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

References

[1] J. Thickstun, Z. Harchaoui and S. M. Kakade, "Learning Features of Music from Scratch," in *International Conference on Learning Representations (ICLR)*, 2017.

[2] Y. N. Hung and Y. H. Yang, "Frame-level Instrument Recognition by Timbre and Pitch," in *Proceedings of the International Society of Music Information Retrieval (ISMIR)*, Paris, 2018.

[3] Y. Han, J. Kim and K. Lee, "Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, pp. 208-221, 1 2017.

[4] P. Li, J. Qian and T. Wang, "Automatic Instrument Recognition in Polyphonic Music Using Convolutional Neural Networks," *arXiv e-prints,* p. arXiv:1511.05520, 11 2015.

[5] A. A. Nugraha, A. Liutkus and E. Vincent, "Multichannel Audio Source Separation With Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, pp. 1652-1664, 9 2016.

[6] P. Chandna, M. Miron, J. Janer and E. Gómez, "Monoaural Audio Source Separation Using Deep Convolutional Neural Networks," in *Proceedings of the International Conference of Latent Variable Analysis and Signal Separation (LVA-ICA)*, Grenoble, 2017.

[7] G. Roma, O. Green and P. A. Tremblay, "Improving Single-Network Single-Channel Separation of Musical Audio with Convolutional Layers," in *Proceedings of the International Conference of Latent Variable Analysis and Signal Separation (LVA-ICA*, Guildford, 2018.

[8] D. Stoller, S. Ewert and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network For End-to-End Audio Source Separation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, 2018.

[9] J. Salamon, E. Gómez, D. P. W. Ellis and G. Richard, "Melody Extraction from Polyphonic Music Signals: Approaches, Applications and Challenges," *IEEE Signal Processing Magazine,* vol. 31, pp. 118-134, 3 2014.

[10] Z. Rafii, A. Liutkus, F. R. Stöter, S. I. Mimilakis, D. FitzGerald and B. Pardo, "An Overview of Lead and Accompaniment Separation in Music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 26, pp. 1307-1335, 8 2018.

[11] P. Huang, M. Kim, M. Hasegawa-Johnson and P. Smaragdis, "Deep learning for monaural speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014.

[12] N. Takahashi, N. Goswami and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," *arXiv e-prints,* p. arXiv:1805.02410, 5 2018.

[13] 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.

[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.

7.3 QMUL3 "Leveraging user interaction to learn performance tracking"

References

[1] Simon Dixon. Live tracking of musical performances using on-line time warping. Citeseer.

[2] Matthias Dorfer, Andreas Arzt, and Gerhard Widmer. Towards score following in sheet music images. arXiv preprint arXiv:1612.05050, 2016.

[3] Taegyun Kwon, Dasaem Jeong, and Juhan Nam. Audio-to-score alignment of piano music using rnn-based automatic music transcription. arXiv preprint arXiv:1711.04480, 2017.

[4] Matthias Dorfer, Jan Hajic Jr, Andreas Arzt, Harald Frostel, and Gerhard Widmer. Learning audio/sheet music correspondences for cross-modal retrieval and piece identification. Transactions of the International Society for Music Information Retrieval, 1(1), 2018.

[5] Matthias Dorfer, Florian Henkel, and Gerhard Widmer. Learning to listen, read, and follow: Score following as a reinforcement learning game. arXiv preprint arXiv:1807.06391, 2018.

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

References

[1] Stowell, Dan, and Mark D. Plumbley. "Delayed decision-making in real-time beatbox percussion classification." *Journal of New Music Research* 39.3 (2010): 203-213.

[2] Ramires, António, Rui Penha, and Matthew EP Davies. "User Specific Adaptation in Automatic Transcription of Vocalised Percussion." *arXiv preprint arXiv:1811.02406* (2018).

[3] Mehrabi, Adib, Kuenwoo Choi, Simon Dixon and Mark Sandler. "Similarity measures for vocal-based drum sample retrieval using deep convolutional auto-encoders." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[4] Alejandro Delgado, SKoT McDonald, Ning Xu, and Mark Sandler. 2019. "A New Dataset for Amateur Vocal Percussion Analysis." *Audio Mostly (AM'19)*, Nottingham, United Kingdom.

[5] Cartwright, Mark, and Bryan Pardo. "Vocalsketch: Vocally imitating audio concepts." *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

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

References

[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).

[4] Schlüter, J., & Grill, T. (2015). Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks. International Society of Music Information Retrieval (ISMIR).

7.6 UPF1 "Facilitating Interactive Music Exploration"

[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.

References

[1] Cohen, W. W. & Fan, W. Web-collaborative filtering: recommending music by crawling the web. Computer Networks 33, 685{698 (2000). URL https://doi.org/10.1016/S1389-1286(00)00057-8.

[2] Loeb, S. Architecting personal delivery of multimedia information. Commun. ACM 35, 39{48 (1992). URL https://doi.org/10.1145/138859. 138862.

[3] Nam, J., Choi, K., Lee, J., Chou, S. & Yang, Y. Deep learning for audio- based music classification and tagging: Teaching computers to distinguish rock from bach. IEEE Signal Process. Mag. 36, 41{51 (2019). URL https://doi.org/10.1109/MSP.2018.2874383.

[4] Choi, K., Fazekas, G. & Sandler, M. B. Automatic tagging using deep convolutional neural networks. In Mandel, M. I., Devaney, J., Turnbull, D. & Tzanetakis, G. (eds.) Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016. URL https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/009_Paper.pdf.

[5] Law, E., West, K., Mandel, M. I., Bay, M. & Downie, J. S. Evaluation of algorithms using games: The case of music tagging. In Hirata, K., Tzanetakis, G. & Yoshii, K. (eds.) Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, Kobe International Conference Center, Kobe, Japan, October 26-30, 2009, 387-392 (International Society for Music Information Retrieval, 2009). URL http://ismir2009.ismir.net/proceedings/OS5-5.pdf.

[6] Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B. & Lamere, P. The million-song dataset. In Klapuri, A. & Leider, C. (eds.) Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011, 591{596 (University of Miami, 2011). URL http://ismir2011.ismir.net/papers/OS6-1.pdf.

[7] Defferrard, M., Benzi, K., Vandergheynst, P. & Bresson, X. FMA: A dataset for music analysis. In Cunningham, S. J., Duan, Z., Hu, X. & Turnbull, D. (eds.) Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017, 316{323 (2017). URL https://ismir2017.smcnus.org/wp-content/uploads/2017/10/75_Paper.pdf.

[8] 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.

[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.

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

References

[1] Masataka Goto, "RWC music database: Popular, classical, and jazz music databases," in 3rd International Society for Music Information Retrieval Conference (ISMIR), 2002, pp.287–288.

[2] Oriol Romani Picas, Hector Parra Rodriguez, Dara Dabiri, Hiroshi Tokuda, Wataru Hariya, Koji Oishi, and Xavier Serra, "A real-time system for measuring sound goodness in instrumental sounds," in Audio Engineering Society Convention 138, Warsaw, Poland, 2015, p. 9350.

[3] Jordi Pons, Olga Slizovskaia, Rong Gong, Emilia Gómez, and Xavier Serra, "Timbre analysis of music audio signals with convolutional neural networks," in 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017, pp. 2744–2748

[4] Yoonchang Han, Jaehun Kim, Kyogu Lee, Yoonchang Han, Jaehun Kim, and Kyogu Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 25, no. 1, pp. 208–221, Jan. 2017.

[5] Peter Li, Jiyuan Qian, and Tian Wang, "Automatic instrument recognition in polyphonic music using convolutional neural networks," arXiv preprint arXiv:1511.05520, 2015.

[6] Brian McFee, Eric J Humphrey, and Juan Pablo Bello, "A software framework for musical data augmentation.," in 16th International Society for Music Information Retrieval Conference (ISMIR), 2015, pp. 248–254.

[7] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Processing Letters, vol.24, no. 3, pp. 279–283, 2017.

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

References

[1] Lee, J., Chang, S., Choe, S. K. & Lee, K. Cover song identification using song-to-song cross-similarity matrix with convolutional neural network. In Proc. of 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 396–400 (Calgary, AB, Canada, 2018).

[2] Xu, X., Chen, X. & Yang, D. Key-invariant convolutional neural network toward efficient cover song identification. In Proc. of 2018 IEEE Int. Conf. on Multimedia and Expo (ICME), 1–6 (San Diego, CA, USA, 2018).

[3] Qi, X., Yang, D. & Chen, X. Triplet convolutional network for music version identification. In Schoeffmann, K. et al. (eds.) Multimedia Modeling, 544–555 (Springer Int. Publishing, 2018).

[4] Doras, G. & Peeters, G. Cover detection using dominant melody embeddings. In Proc. of 20th Int. Conf. on Music Information Retrieval (ISMIR) (Delft, The Netherlands, 2019).

[5] Ye, Z., Choi, J. & Friedland, G. Supervised deep hashing for highly efficient cover song detection. In Proc. of 2019 IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR), 234–239 (San Jose, CA, USA, 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.

7.9 TPT1 "Behavioral music data analytics"

References

[1] Jesus Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutierrez. Recommender systems survey. *Knowledge-based systems*, 46:109–132, 2013

[2] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: introduction and challenges. In *Recommender systems handbook*, pages 1–34. Springer, 2015.

[3] G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, June 2005.

[4] Jean-julien Aucouturier and Francois Pachet. Music Similarity Measures: What is the Use. In Proceedings of the *ISMIR*, pages 157–163, 2002.

[5] Beth Logan. Music Recommendation from Song Sets. In *International Conference on Music Information Retrieval* 2004, number October, pages 10–14, Barcelona, Spain, 2004.

[6] Chun-man Mak, Tan Lee, Suman Senapati, Yuting Yeung, and Wang-kong Lam. Similarity Measures for Chinese Pop Music Based on Low-level Audio Signal Attributes. In 11th International Society for Music Information Retrieval Conference, number *ISMIR*, pages 513–518, 2010

[7] Yoo, So-Hyeon, Seong-Woo Woo, and Zafar Amad. "Classification of three categories from prefrontal cortex using LSTM networks: fNIRS study." 2018 18th International Conference on Control, Automation and Systems (ICCAS). IEEE, 2018.

[8] Pons, Jordi, Thomas Lidy, and Xavier Serra. "Experimenting with musically motivated convolutional neural networks." *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2016.

7.10 TPT2 "Voice models for lead vocal extraction and lyrics alignment"

References

[1] 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.

[2] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde. Singing Voice Separation with Deep U-Net Convolutional Networks. In Proceedings of the International Society for Music Information Retrieval Conference, pages 745–751, 2017.

[3] N. Takahashi and Y. Mitsufuji. Multi-scale multi-band densenets for audio source separation. In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pages 21–25. IEEE, 2017.

[4] D. Stoller, S. Ewert, and S. Dixon. Jointly detecting and separating singing voice: A multi-task approach. In International Conference on Latent Variable Analysis and Signal Separation, pages 329–339. Springer, 2018.

[5] D. Stoller, S. Ewert, and S. Dixon. Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation. In Proceedings of the International Society for Music Information Retrieval Conference, pages 334–340, 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.

[6] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. In SSW, page 125, 2016.

[7] F. Lluís, J. Pons, and X. Serra. End-to-end music source separation: is it possible in the waveform domain? arXiv preprint arXiv:1810.12187, 2018.

[8] D. Rethage, J. Pons, and X. Serra. A wavenet for speech denoising. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 5069–5073. IEEE, 2018.

[9] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner. The musdb18 corpus for music separation, 2017.

7.11 TPT3 "Multimodal movie music track remastering"

References

[1] B. Kaneshiro, D. T. Nguyen, J. P. Dmochowski, A. M. Norcia, and J. Berger, "Naturalistic music eeg dataset - hindi (nmed-h)," https://purl.stanford.edu/sd922db3535, 2016.

[2] S. Losorelli, D. T. Nguyen, J. P. Dmochowski, and B. Kaneshiro, "Nmed-t: A tempo-focused dataset of cortical and behavioral responses to naturalistic music," 2017.

[3] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Towards music imagery information retrieval: Introducing the openmiir dataset of eeg recordings from music perception and imagination." in ISMIR, 2015, pp. 763–769

[4] M. S. Treder, H. Purwins, D. Miklody, I. Sturm, and B. Blankertz, "Decoding auditory attention to instruments in polyphonic music using single-trial eeg classification," Journal of neural engineering, vol. 11, no. 2, p. 026009, 2014.

7.12 TPT4 "Context-driven music transformation"

References

[1] Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou and Alexei A. Efros. "Image-to-Image Translation with Conditional Adversarial Networks." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 5967-5976.

[2] Sangkloy, Patsorn, Jingwan Lu, Chen Fang, Fisher Yu and James Hays. "Scribbler: Controlling Deep Image Synthesis with Sketch and Color." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 6836-6845.

[3] Zhu, Jun-Yan, Taesung Park, Phillip Isola and Alexei A. Efros. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks." 2017 IEEE International Conference on Computer Vision (ICCV) (2017): 2242-2251.

[4] Liu, Ming-Yu, Thomas Breuel and Jan Kautz. "Unsupervised Image-to-Image Translation Networks." ArXiv abs/1703.00848 (2017).

[5] Lample, Guillaume, Alexis Conneau, Ludovic Denoyer and Marc'Aurelio Ranzato. "Unsupervised Machine Translation Using Monolingual Corpora Only." ArXiv abs/1711.00043 (2017).

[6] Mor, Noam, Lior Wolf, Adam Polyak and Yaniv Taigman. "A Universal Music Translation Network." ArXiv abs/1805.07848 (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] Lu, Wei Tsung and Li Su. “Transferring the Style of Homophonic Music Using Recurrent Neural Networks and Autoregressive Model.” ISMIR (2018).

[8] Brunner, Gino, Andres Konrad, Yuyi Wang and Roger Wattenhofer. “MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer.” ArXiv abs/1809.07600 (2018).

[9] Brunner, Gino, Yuyi Wang, Roger Wattenhofer and Sumu Zhao. “Symbolic Music Genre Transfer with CycleGAN.” 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI) (2018): 786-793.

[10] Grinstein, Eric, Ngoc Q. K. Duong, Alexey Ozerov and Patrick Pérez. “Audio Style Transfer.” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017): 586-590.

7.13 TPT5 “Conditional generation of audio using neural networks and its application to music production”

References

[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] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. “Enabling factorized piano music modeling and generation with the maestro dataset.” arXiv preprint arXiv:1810.12247, 2018.

[3] Law, Edith and Von Ahn, Luis. “Input-agreement: a new mechanism for collecting data using human computation games.” *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 1197–1206. ACM, 2009.

[4] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. “Neural audio synthesis of musical notes with wavenet autoencoders”. *In Proceedings of the 34th International Conference on Machine Learning*, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, pages 1068–1077, 2017.

[2] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stoter, Stylianos Ioannis Mimilakis,and Rachel Bittner. “The MUSDB18 corpus for music separation”, December 2017

[3] M Vinyes. MTG MASS database. http://www.mtg.upf.edu/static/mass/resources, 2008.

[4] TELEFUNKEN Elektroakustik, Studio sessions. https://telefunken-elektroakustik.com/multitracks, 2001.

[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.

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

References

[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.

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

References

[1] Dannenberg, R. B. (1984). 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*.

[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.

[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).

[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)*.

[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.