New Frontiers in Music Information Processing (MIP-Frontiers)

Grant Agreement Number: 765068

• Title:	State of the art, challenges and potential of user-driven approaches in MIR
• Lead Beneficiary:	TPT (Télécom Paris)
• 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 user-driven aspects – the focus of the present report – the main message is that MIR research needs to include the user perspective, in order to produce artefacts which are relevant and have an impact on their market. It is therefore essential to understand the user's role within the music communication chain and to develop user-centred interaction-based technologies.

The preparation of this report started with an internal discussion, within the consortium, on what user-driven methods can mean for MIR in general, and for the PhD students' projects in particular. It was obvious that all of the 15 MIP-Frontiers projects targets aspects of user-driven methods, some directly and others implicitly in their 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 user-driven methods, such as music as a cultural and social phenomenon that involves directly the listener, identifying and using the relevant user contexts, and/or transferring the style from music coming from a similar context. This report has already been, and will be, a valuable resource, demonstrating to the ESRs and to the MIR community the importance of user-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 user-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 "user-driven methods and approaches" is understood in the framework of MIP-Frontiers. As the MIP-Frontiers Training Network comprises 15 individual ESRs with different topics in the MIR field, we obtain fifteen different views on relevant scientific background, challenges, opportunities, and solutions.

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

3. Introduction

In the project proposal, the MIP-Frontiers consortium identified three big challenges – and thus, from a scientific point of view, research opportunities – for the further development of the MIR field. One of these is the need for more user-driven aspects in MIR systems, and in the research and development process (in addition to large amounts of data). Specifically, this was argued as follows: *The user perspective is a core aspect of MIR, as music is a cultural phenomenon in which users are central. In order to produce artefacts which are relevant and have an impact on their market, it is therefore essential to understand user roles within the music communication chain, to evaluate the MIR technologies from the perspective of the user (i.e. in their actual use context), and to develop user-centred interaction-based technologies. While much MIR work considers the user only implicitly, e.g. as a source of ground truth, we believe it is paramount for MIR researchers and system designers to define and implement user-centred methodologies in creating their musical applications.*

Correspondingly, the objectives of user-driven approaches in MIR, as defined in the project web, https://mip-frontiers.eu/, and in the project proposal, are to better considering the user in MIR systems: in informed source separation where the user can provide specific information to the system to guide the separation; in interactive systems between user and machine; and in exploiting the user's local context, behaviour or mood in music recommendation systems.

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

4. User-driven methods in MIP-Frontiers

4.1 The concept of the user in MIR

Because music is (usually) created by humans for humans, the MIR domain has a special need to place the user at the heart of its research. Obviously, data-driven and knowledge-driven methods carry and exploit information about usage or user context, but this is in most cases in an implicit form. Better considering the user in MIR is therefore an essential research dimension.

In the context of the present project, we will mostly refer to the work that either exploits direct user data while processing an MIR task, or that builds a system which exploits user information such as context with or without the user's explicit cooperation.

The rest of this section lists those projects in MIP-Frontiers where user-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 user-driven aspects

QMUL1: “Representation Learning in Singing Voice”

In this project, the specific goal is to cluster user data using unsupervised learning methods. The dataset provided by DoReMir is well-suited for the task due to its size and high variance. The clusters will be analyzed to observe any contextual information. Through clustering, the project aims at creating clean and refined training datasets.

QMUL2: “Improving Polyphonic Transcription through Instrument Recognition and Source Separation”

Project QMUL2 focus on better understanding the aspects and qualities of music sounds that are related to the timbre of musical notes and that forces 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 proper symbols can be applied in the transcription to represent them.

In this project, there will not be any exploitation of the user's local context, behaviour or mood. The only user-driven method that can be related to it is the potential user interaction with the system in order to allow them to choose which instruments they want to be extracted from the recording or which instrumental line they need to have transcribed.

QMUL3: “Leveraging user interaction to learn performance tracking”

QMUL3 is focused on music alignment. Music alignment aims at providing a way to navigate among various music representations in a unified manner, lending itself applicable to a myriad of domains like music education, performance, enhanced listening, automatic accompaniment and so on. This project has a significant user-driven component, since it involves exploiting the information obtained from the users to improve music alignment. We also wish to advance user-centric MIR research through this project since work in the field of Music Information Retrieval (MIR) has been largely systems-focused despite recurring calls for a greater focus on user-centric research.

QMUL4: “Robust Timbre Analysis for Query by Vocal Imitation”

The process of querying by vocal imitation does not necessarily have to end when the most similar sound in the dataset is retrieved. As both our vocal apparatus and our skills to control it are significantly limited when we want to imitate complex sounds, the retrieval system will often output an unsatisfying result to the user. There are different strategies to refine the search result in order to finally get the sound that the user was imitating, the most effective ones being those that directly involve the users themselves.

[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.

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. The user-driven aspect is related to the data- and knowledge-driven technologies, and is described in those approaches.

TPT1: “Behavioral music data analytics”

Project TPT1 aims at improving music recommendation systems by integrating users’ contextual information in the recommendation process. Users’ contextual information is defined as the external factors that affect their music preferences at any given time. For example, user activity, location or time of the day are considered as contextual information that change the user’s preferences. For certain activities the user would prefer to listen to energetic music while in some others he/she would prefer to listen to calming music. Hence, we need to consider the user’s preferences, common contexts, and the influence of these contexts on the user’s preferences.

TPT3: “Multimodal movie music track remastering”

Project TPT3 aims at developing new methods for offering the user an interactive multimedia experience. In particular, we want to develop methods for a user-centered remastering of music recordings. The idea is to guide audio source separation/enhancement using the user’s attention as a high-level control/feedback to select which, for him/her, is the desired source to enhance.

TPT4*:* “Context-driven music transformation”

The aim of project TPT4 is to enable transforming music in terms of artistic style. A possible application is to adapt music to the needs of a specific user, either based on examples of the user's taste, or on other user-dependent data (e.g., location, mood).

TPT5: “Conditional generation of audio using Neural Networks and its application to Music Production”

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

UPF1: “Facilitating Interactive Music Exploration”

Project UPF1 aims to introduce a user-centric music exploration system that will take advantage of user feedback and interactions to personalize and optimize the exploration process with the help of reinforcement 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.

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. An evaluation with music makers on the different algorithms for loop characterization will be conducted to understand how each algorithm is useful for this task.

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. Although we hope that those systems would facilitate finding different versions of favorite songs for the end users, this project mainly focuses on data-driven and knowledge-driven approaches. For more information, we refer the reader to the respective deliverables.

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

Project JKU1 aims at developing new algorithms for the automatic structuring and cross-linking of large multi-modal music collections, with a focus on audio recordings and sheet music images (acoustic and visual domains, respectively). In addition to 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, comprising audio recordings and score images in various representations. Although we believe that those algorithms would allow users to explore large music collections in a convenient and enjoyable way, the main focus of this project is on knowledge-driven and data-driven approaches, for which the reader is suggested to go through the corresponding deliverable documents.

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. Although we expect the user to benefit from the technology to be developed in this project to guide them through a musical work, the proposed research does not focus on the user perspective but rather on musical knowledge and data useful to the task. For more information, we refer the reader to the respective deliverables.

[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 State of the Art, Challenges, and Opportunities (by Project)

5.1 QMUL1 "Representation Learning in Singing Voice"

5.1.1 Project Goals

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 from real-world user data and further create clean and refined subsets from big unlabelled datasets for the purpose of training supervised learning algorithms.

5.1.2 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 a 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 representations describing the sequential structures of the audio segments. The authors of the paper [3] propose 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.

5.1.3 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) [4-7] also provide great potential for the specific task.

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

5.2.1 Project Goals

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

5.2.2 State of the Art

There is no state-of-the-art user-driven methods related to this project. For data or knowledge driven methods, please check the respective deliverables.

[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.

5.2.3 Challenges and Opportunities

The only user-driven opportunity that is related to this project is to allow the user choose which instruments they want to be extracted from the recording or which instruments they need to have transcribed, and thereof having its notes recognised.

5.3 **QMUL3** “Leveraging user interaction to learn performance tracking”

5.3.1 Project Goals

The project QMUL3 is aimed at developing robust music alignment techniques. The resulting alignment model is thought to have the following properties:

- is general enough to cater to different acoustic settings and instruments,
- at the same time can adapt to the setting in which it is being employed,
- does not require complex feature engineering on the manual front,
- at the same time is able to leverage human feedback and improve itself.

These properties will be achieved by employing user-centric deep learning methods.

5.3.2 State of the Art

The extraction of music information from audio has been studied to a considerable extent in recent work in MIP, however the consideration of the performance environment and the user perception component still remains largely unexplored [1-4]. There have been some recent works which incorporate user-centric factors in MIR. [5] leverages user input and feedback for Interactive Sound Event Detection and Annotation. [6] shows that adapting to a specific piano yields about 10% improvement in the accuracy of an automatic transcription system. [7] propose an XML-like language, UMIRL (User Modeling for Information Retrieval Language), in which different systems may describe the user in this standard format to make the user model sharable and reusable. [1] and [8] show the effectiveness of incorporating contextual information from the user to improve MIR tasks, with the former focussing on alignment.

5.3.3 Challenges and Opportunities

A major challenge in this direction is the gathering of user data - this can be challenging depending upon the level of the sensitivity of the information to be collected as well as the concerns of the user. Another challenge is to ensure the quality of the said collected data, since it is not manually labelled by experts. Another challenge is designing the interactions and the way in which the data would be used by the model.

We will investigate the performance of music alignment which uses an “online” learning approach in which, at test stage, the model is continuously adapted to a stream of incoming alignment corrections. In machine learning, online learning is defined as the task of using data that becomes available in a sequential order to step-wise update a predictor for future data. The new data becoming available in our case would be the incoming alignment corrections. These corrections would ideally be coming from the user but we can also explore methods where we could generate these corrections (semi-)automatically. Thus, we will use online learning to perform the exploitation of manual alignment corrections in a continuous learning framework.

[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.

We could explore an automatic alignment correction approach in resource-scarce conditions, especially when generating robust offline alignments is preferable - for instance, for quick deployment on an iPad. This can be thought of more as a domain adaptation process. Continuous learning is ideal in a resource-intensive scenario, where we have a constant flux of new incoming labels. Our model would constantly improve itself using active learning and online learning strategies.

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

5.4.1 Project Goals

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.

5.4.2 State of the Art

Two algorithmic routines have been proposed in query by sound example that let the users refine their original query. SynthAssist [9] is a system where the user is presented with similar sounds to the original vocal imitation and is asked to rate them in perceptual closeness with the imitated (target) sound. Based on the responses, the sound content of the original vocal imitation is updated (transformed) and this new file is fed back to the algorithm. This generates a new set of similar sounds for the user to rate until the imitated sound is finally found.

The system proposed in [10] works in a similar way to SynthAssist, but the update is a new vocal imitation given by the user, which ideally embodies how the user would want the suggested similar sounds to sound like.

5.4.3 Challenges and Opportunities

We believe that deep learning-based timbral transformations can help existing user-driven models to converge faster to the imitated sound. Also, as different users have a distinct way of imitating sounds, a good idea would be to personalise these routines by making a model of the user's vocal imitation style. Using robust timbre analysis along with efficient interactive learning techniques like active learning [11] appears to be a promising way to deal with users' idiosyncrasies in this respect.

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

5.5.1 Project Goals

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.

5.5.2 State of the Art

No user-driven techniques are related to the 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.

5.5.3 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. As described before, no user-driven specific challenge will be addressed.

5.6 **TPT1** "Behavioral music data analytics"

5.6.1 Project Goals

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

5.6.2 State of the Art

One of recent and relevant studies is a PhD thesis by Martin Pichl [12]. In summary, the thesis focused on using contextual information in music recommendation through playlist names [13] along with audio content, user demographics and social media activity on Twitter [14]. The study included the development of a suitable dataset along with the evaluation of different models (the main proposed model is Factorization Machines [15]). Pichl's work is relevant and showed improvement in recommendation using context. However, it faced challenges in the data available to use. Relying on Twitter to understand users' behavior is somehow limiting because the study focuses only on a subset of users that are actively tweeting their listening events. Additionally, they had limited access to acoustic features using only 7 features available from the Spotify API. In our case, with access to additional content-based data from Deezer, we can use the concept of using the playlist names in defining the context while also making use of audio content which can be described by appropriate acoustic features.

These previous studies approached the problem through a user-driven methodology to automatically identify the user's contexts and build a dataset based on user created playlists. However, they lacked further analysis of the discovered contextual information and the relationship between them. Our preliminary studies showed that most of the clusters used in these experiments were barely contextual clusters and included more information about genres and music style. However, the methodology of deriving context classes using playlist titles is still a relevant way to derive a dataset of context classes using user created data.

5.6.3 Challenges and Opportunities

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

Similarly, there are no available standard datasets for this problem, in terms of tracks being labelled with their context classes. This is another challenge that is important for future research to have a baseline and a dataset to compare new results with. Relying on user-created data such as playlists is a suitable approach to create a semi-automated procedure of collecting and labelling tracks using the 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.

5.7 **TPT3** "Multimodal movie music track remastering"

5.7.1 Project Goals

The goal of the project is to perform 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.

5.7.2 State of the Art

Informed source separation exploits all the available prior information about the sources and the mixing process along with the audio signal [16] and was proven to enhance the source separation process especially for music. Many works have been proposed which involve other modalities such as the score [17], the text [18] and the motion of sound sources [19]. Only a few works have been proposed in the last years that involve the user and in particular that combines source separation/enhancement with auditory attention decoding characterized by EEG recordings [20-22]. However, they all focus on attention to speech stimuli. We aim to take advantage of these recent works in order to develop a new form of informed music source separation approach which can be referred as neuro-steered music source separation. This task, to my knowledge, was never addressed before.

5.7.3 Challenges and Opportunities

In this project, the user has a fundamental role. Music is considered not in an abstract way, but as a cultural and social phenomenon that directly involves the listener. This implies a lot of possible future development of the project in real applications but at the same time implies many problems related to the lack of data. Collecting human data is very expensive and time-consuming, and involves many more variables. This is the main reason why only few and small datasets are available in this field.

5.8 **TPT4** "Context-driven music transformation"

5.8.1 Project goals

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).

5.8.2 State of the art

Precursor works proposed to retrieve music from a database which is well adapted to the video content [23] (in terms of correlation between the music rhythm and movement dynamics in the video scene) or to transform the rhythmic expressiveness of audio recordings [24]. More recent studies aim at retargeting music based on the video content by exploiting automatic music segmentation and rhythm analysis.

More recently, it has been proposed to perform audio style transfer using techniques developed for images [25]. 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.

[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.

5.8.3 Challenges and opportunities

High-quality music style transfer would open the possibility for user-dependent applications by means of transferring the style from music coming from a similar context. However, transforming or generating audio using deep learning techniques (which is a natural choice for this task) is still very challenging and resource intensive. One alternative is to work with music in a symbolic representation, which is more abstract than audio waveforms or spectrograms and, due to its discrete nature, easy to generate using RNN or transformer models.

5.9 **TPT5** "Conditional generation of audio using Neural Networks and its applications to Music Production"

5.9.1 Project Goals

From a user-driven perspective, project TPT5 aims at developing new musically motivated and interactive methods for navigating sonic spaces. In particular, we want to develop methods for controlling audio synthesis in a musically meaningful way, avoiding complicated controls existent in modern synthesizers in favor of rather musically-driven and user-friendly parameters inferred from big data. We approach this task by learning disentangled latent representations of musical relationships by conditioning the generative model on some pre-existent audio content.

5.9.2 State of the Art

Learning representations of data that are meaningful to humans has been a topic of significant interest in recent years. The field of Computer Vision has seen some of the most remarkable breakthroughs, where Neural Networks have been capable of learning, in a completely unsupervised way, stylistic parameters for very abstract features. For instance, GANs can infer independent parameters for hairstyles, face expressions, poses or accessories (e.g., glasses, hats, etc.) on the task of face image generation [26] or, given a simple painted sketch of a landscape, render a high-quality real image version of it [27]. In the field of audio, although behind Computer Vision, the first neural-network-driven synthesizer enables users the blending of instruments, creating entirely new sounds by navigating in a two-dimensional timbre space [28]. Further research has been aimed to impose even more musical structure into the latent space learned by the models [29].

5.9.3 Challenges and Opportunities

Although there have been considerable advances in user-driven approaches for generative models of audio, the trend is that the field lacks behind Computer Vision research. Although this might seem to offer many opportunities, this fact is probably due to the multi-scale complexity and sequential nature of music audio, as opposed to image, which makes the field far more difficult.

[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 **UPF1** "Facilitating Interactive Music Exploration"

5.10.1 Project Goals

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

5.10.2 State of the Art

As we are proposing a user-centric evaluation approach, the metrics that will be used include various user engagement ones [30], entropy-based novelty [31] and other user-centric metrics [32]. Moreover, to optimize the exploration process we use reinforcement learning [33] with users providing the state and reward.

5.10.3 Challenges and Opportunities

User-centric research in MIR had always been more difficult due to the user evaluations which are necessary. However, collaborating with the industrial partner Jamendo in developing the prototype and deploying it in an A/B testing environment provides a great opportunity to focus on user evaluations and use the feedback for improving the system.

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

5.11.1 Project Goals

In order to evaluate how different algorithms for audio characterisation perform on retrieving loops from large-scale audio databases, we want to conduct user studies with music makers. Besides this, we want to develop an interface for loop retrieval in said databases and evaluate if using this system can empower creativity and reduce the time for retrieving a loop which fits the music maker's composition.

5.11.2 State of the Art

The evaluation methodologies in MIR can be divided in system-based and user-centred. As defined in [34], system-based MIR includes all *research concerned with experiments existing solely in a computer*, while user-centered MIR *involves human subjects and their interaction with MIR systems*. Despite recurring calls for a need to include user evaluation in this field, most of the work has largely been focused in system-based MIR [35].

While for automatic classification tasks such as instrument classification, the ground truth is easy to classify, for more perceptual based tasks such as music similarity, a ground truth is harder to define. This topic was deeply studied in [36], which provides a series of valuable insights. In order to compare algorithms for retrieval, the author proposes a experiment design. This task template allows to both evaluate two individual systems and their relative performance to assess which one will provide more user satisfaction. Besides this, the author also suggests an evaluation metric which we see as relevant towards a quantitative analysis of our research which is how long it took for a user to complete the task proposed. In a loop retrieval system, this can be considered as the number of loops which the user had to preview until one that fits their intentions is achieved.

[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.

This evaluation of musical similarity can be useful for comparing algorithms for retrieval. However, in creative tasks, the accuracy of the system should not be the only evaluation metric to be used. In our case, the goal is not to locate specific musical content known a priori, but to discover a loop which fits what the music maker has in mind. As stated in [37], *users are engaging with the technology in order to express themselves through it* and, therefore, metrics for evaluating how the system can help creativity should be employed. In [38], a methodology for evaluating how a system can foster creativity is proposed, entitled the Creativity Support Index. This methodology comprises a survey where several concepts and factors associated to creativity are evaluated in a system.

5.11.3 Challenges and Opportunities

In order to evaluate the retrieval of loops for assisting music production, we have to evaluate how the different music information retrieval algorithms empower creativity and assist music makers on finding a loop which fits their necessities.

These algorithms need a user-based validation process which can be achieved through the community of Freesound, a community audio collection which was developed in Universitat Pompeu Fabra. We want to evaluate these algorithms with user tests, to incorporate them into a loop retrieval system.

Finally, we want to evaluate this loop retrieval system on how it can improve the loop retrieval in terms of how many loops the users had to listen to until they found a relevant one and on how this system can empower creativity, based on the Creativity Support Index.

6 Conclusion

This deliverable summarizes the main aspects linked to the User-driven approaches that are at the heart of the MIP-Frontiers projects. In essence, all projects involve in some sense the user, but in this deliverable we mainly discuss the projects which have a more substantial user-driven aspect. For those projects, we have recalled the project goals and we provided a brief state of the art for each project to highlight in which sense the projects involve user-driven concepts and strategies and how these projects are linked to the current state of the art.

[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 References

[1] «Meng Wang and Xian-Sheng Hua. Active learning in multimedia annotation and retrieval: A survey. ACM Transactions on Intelligent Systems and Technology (TIST), 2(2):10, 2011.».

[2] «Jin Ha Lee and Sally Jo Cunningham. Toward an understanding of the history and impact of user studies in music information retrieval. Journal of Intelligent Information Systems, 41(3):499–521, 2013.».

[3] «Markus Schedl, Arthur Flexer, and Juli´an Urbano. The neglected user in music information retrieval research. Journal of Intelligent Information Systems, 41(3):523–539, 2013.».

[4] «Gerhard Widmer. Getting closer to the essence of music: The con espressione manifesto. ACM Transactions on Intelligent Systems and Technology (TIST), 8(2):19, 2017.».

[5] «Bongjun Kim. Leveraging user input and feedback for interactive sound event detection and annotation. In 23rd International Conference on Intelligent User Interfaces, pages 671–672. ACM, 2018.».

[6] «Sebastian Ewert and Mark Sandler. Piano transcription in the studio using an extensible alternating directions framework. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(11):1983–1997, 2016.».

[7] «Wei Chai and Barry Vercoe. Using user models in music information retrieval systems. In ISMIR, 2000.».

[8] «Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. In Advances in neural information processing systems, pages 2643–2651, 2013.».

[9] «Cartwright, Mark, and Bryan Pardo. "Synthassist: an audio synthesizer programmed with vocal imitation." Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014.».

[10] «Kim, Bongjun, and Bryan Pardo. "Improving Content-based Audio Retrieval by Vocal Imitation Feedback." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.».

[11] «Tong, Simon. Active learning: theory and applications. Vol. 1. USA: Stanford University, 2001.».

[12] «Martin Pichl.Multi-Context-Aware Recommender Systems: A Study on Music Recommendation. PhD thesis, University of Innsbruck, 2018.».

[13] «Martin Pichl, Eva Zangerle. Towards a Context-Aware Music Recommendation Approach: What is Hidden in the Playlist Name? In2015 IEEE International Conference on Data Mining Workshop (ICDMW), pages 1360–1365, November 2015.».

[14] «Eva Zangerle, Martin Pichl, Wolfgang Gassler, and Gnther Specht. #Nowplaying music dataset: Extracting listening behavior from twitter. In Proceedings of the 1st ACMInternational Workshop on Internet-Scale Multimedia Management, ISMM '14, pages21–26, Orla».

[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.

[15] «Steffen Rendle. Factorization Machines. In 2010 IEEE International Conference on Data Mining, pages 995–1000, December 2010.».

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

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

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

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

[20] «N. Das, S. Van Eyndhoven, T. Francart, and A. Bertrand. “Eeg-based attention-driven speech enhancement for noisy speech mixtures using n-fold multi-channel wiener filters”. In 25th European Signal Processing Conference (EUSIPCO), pages 1660–1664. IEEE, 20».

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

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

[23] «Gillet, Olivier, Slim Essid and Gaël Richard. “On the Correlation of Automatic Audio and Visual Segmentations of Music Videos.” IEEE Transactions on Circuits and Systems for Video Technology 17 (2007): 347-355.».

[24] «Gouyon, Fabien, Lars Fabig and Jordi Bonada. “Rhythmic Expressiveness Transformations Of Audio Recordings: Swing Modifications.” (2003).».

[25] « 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.».

[26] «Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks.” CoRR, abs/1812.04948, 2018.».

[27] «Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.».

[28] «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 Learnin».

[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.

[29] «Philippe Esling, Axel Chemla-Romeu-Santos, and Adrien Bitton. “Bridging audio analysis, perception and synthesis with perceptually-regularized variational timbrespaces.” In Proceedings of the 19th International Society for Music Information Retriev».

[30] «Lehmann, J., Lalmas, M., Yom-Tov, E. & Dupret, G. Models of user en-gagement. In Masthoff, J., Mobasher, B., Desmarais, M. C. & Nkambou, R.(eds.)User Modeling, Adaptation, and Personalization - 20th InternationalConference, UMAP 2012, Montreal, Canada, J».

[31] «Bellog ́ın, A., Cantador, I. & Castells, P. A study of heterogeneity in rec-ommendations for a social music service. InProceedings of the 1st In-ternational Workshop on Information Heterogeneity and Fusion in Recom-mender Systems, HetRec ’10, 1–8».

[32] «Celma,`O. & Herrera, P. A new approach to evaluating novel recommen-dations. In Pu, P., Bridge, D. G., Mobasher, B. & Ricci, F. (eds.)Proceed-ings of the 2008 ACM Conference on Recommender Systems, RecSys 2008,Lausanne, Switzerland, October 23-25, 2008,».

[33] «Sutton, R. S., Barto, A. G.et al. Introduction to reinforcement learning,vol. 2 (MIT press Cambridge, 1998). URLhttps://mitpress.mit.edu/books/reinforcement-learning-second-edition».

[34] «Schedl, Markus, and Arthur Flexer. "Putting the User in the Center of Music Information Retrieval." In Proceedings of the 13th International Society for Music Information Retrieval Conference, 2012.».

[35] «Weigl, David M., and Catherine Guastavino. "User studies in the Music Information Retrieval Literature." In Proceedings of the 12th International Society for Music Information Retrieval Conference, 2011.».

[36] «Julian Urbano, “Evaluation in Audio Music Similarity”, PhD dissertation, 2013.».

[37] «Kristina Andersen and Peter Knees. “Conversations with expert users in music retrieval and research challenges for creative MIR”. In Proceedings of the 17th International Society for Music Information Retrieval Conference, 2016.».

[38] «Cherry, Erin & Latulipe, Celine “The creativity support index”. In Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Extended Abstracts Volume, Boston, MA, USA, April 4-9, 2009.».

[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.