New Frontiers in Music Information Processing (MIP-Frontiers)

Grant Agreement Number: 765068

- Title: First report on novel user-driven approaches in MIR
- Lead Beneficiary: TPT
- 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.

1. Introduction

In the original 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 WP3, as defined in Annex 1 to the project agreement, are to better consider 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 deliverable is to provide a first report on this work. It follows the previous deliverable D3.1 which documented the state of the art, current challenges, and corresponding potential of user-driven research approaches to MIR problems. Since "user-driven methods in MIR" is a broad concept, this report will focus on user-related topics as they emerge in the specific research projects (PhD theses) tackled in MIP-Frontiers. The report is then structured by individual PhD projects.

2. User-driven Methods in MIP-Frontiers

The concept of user in MIR

Because music is 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 a MIR task or that builds a system which exploits user information such as context with or without user explicit cooperation

The next 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.

[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. Novel user-driven approaches in MIR (by PhD Project)

3.1 QMUL1: Automatic Lyrics Transcription

It has only been few years that Automatic Lyrics Transcription (ALT) technologies have started to gain interest from various agents including research and industry. The reasons for the recent interest are the recent advances in Deep Learning (and specifically Automatic Speech Recognition - ASR) research and the availability of new data sets. However, there has not been a so-called `Automatic Lyrics Transcription system published or released prior to this research that could reach similar performance or robustness levels with that of ASR. Therefore, we set the end-goal of our research project to develop a robust ALT system that could be leveraged for scientific and commercial music applications. As the methodology, we analyse differences in acoustic, linguistic and pronunciation between singing and speech, and develop singing-adapted models.

In this section we report the novel contributions on the user-driven aspects of this project:

- *The research community*: For scientific usage of the software developed, the proposed system is planned to be shared with the research community having a compact and easy installation procedure and a well-preserved documentation. We plan to exploit Github for storing the code, Docker containers for easy installation and a form of a permalink as pointers to the relevant publication. Furthermore, we plan to develop a vamp-plugin for our system to be used in open-source analysis tools like Sonic Visualiser [1].

- *Mobile applications*: Our industrial collaborator, Doremir Music Research AB designs software for mobile applications. In order to integrate the outcome of this research with an existing mobile application, several steps have to be taken: The research software has to be optimized in terms of size, memory and performance time. This requires a compact version of the research framework and the Automatic Lyrics Transcription (ALT) model, and an efficient way of processing the audio. The installation of the research framework utilized, Kaldi [2], is compressed keeping only the libraries and binaries that are used in feature extraction, transcription and alignment processes. This procedure resulted in shrinking the size of Kaldi from **15.3 GB** to around **1.5 GB.** For the work presented in [3], we employ 'the singer-adaptive' training approach using i-Vectors [4] in the feature space. We use i-Vectors with 100 dimensions, which is 2.5 times higher than that of the acoustic features. Empirical studies showed that i-Vectors are not extremely beneficial for inference cases where singer identity is not explicit. Therefore, we decided to exclude i-Vectors from the feature space and retrain another ALT model. This resulted in reducing the size of the model from **2.5GB** to **1.9GB** with a cost of around 1% Word Error Rate (WER) in the evaluation set used [3]. Moreover, excluding i-Vectors during feature extraction would save some additional time during inference. The language model was initially built using the SRILM Toolkit [5] which has a size of **223MB**. Composing it with the acoustic model resulted in the final ALT model having the size of 1.9GB (without i-Vectors). To reduce the size of the language model, we have trained another model using built-in Kaldi tools that resulted in a language model with **23MB** in size and the final ALT model has **213MB**. This size reduction comes with an additional WER cost of around 1-2%.

- During development, we consider our ALT system to be applicable for streaming.

- Finally, we investigate efficient ways to apply the ALT technology in multiple languages.

[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.

3.2 QMUL 2: Improving Polyphonic Transcription through Instrument Recognition and Source Separation

Automatically transcribing polyphonic music (i.e., with multiple notes playing simultaneously) to a score is a difficult task and one of the most discussed topics in the MIR community. The main challenge comes from the high correlation of sounds that can be played at the same time causing a highly structured overlap of harmonics in the final acoustic signal, which leads to errors in multi-pitch estimation and instrument tracking algorithms. In particular, when analyzing recordings with multiple instruments, the transcription process becomes even more complex, because not only each note should have its pitch and duration properly estimated, but the information regarding the timbre of sounds should also be correctly processed. It is mandatory to have a way of recognizing the instrument that played each note and associate each sound to the correct voice in the final staff notation.

The project QMUL2 proposes to utilise **source separation** and **instrument recognition** techniques with the final goal of improving multi-instrument automatic polyphonic transcription, where not only harmonic, but also percussive instruments may be playing. We started the research focusing on each of those two tasks separately and so far, we have proposed new algorithms to both source separation and instrument recognition task. We pinpoint below user-driven approaches developed in the project during the course of the PhD and we also share our ideas and research directions for the final year of the programme.

Project QMUL2 is a result of a partnership between DoReMIR Music Research and Queen Mary University of London. DoReMIR focuses on developing new software and applications to change the way we play and express ourselves in music. One of its most revolutionary products is ScoreCloud, a software that automatically transcribes music signals directly to a score. However, the two most requested features by its users are the possibility to make the system able to process drum-based signals and perform the transcription along with instrument recognition, being able to assign different instrumental sounds to the correct voice in the final score. In this context, it is expected that the algorithms proposed during the PhD programme will not only benefit the general Music Information Retrieval (MIR) community, but will also generate a new product or be incorporated to ScoreCloud software, which will provide its users more flexibility with a better transcription system.

In the near future we are also planning on doing a study with ScoreCloud users on whether the proposed algorithms actually have improved the automatic music transcription performance of the software, as assessed by actual users instead of only relying on objective evaluation. Therefore, it is possible to say that the users' opinion will have a direct impact on our approach to the task of music transcription. Another user-driven potential application that we see in this project is allowing the users to choose which instruments they want to be extracted from the recording or which instruments they need to have transcribed.

3.3 QMUL3: Leveraging User Interaction to Learn Performance Tracking

QMUL3 is aimed at developing music alignment methods, which can potentially incorporate user information. Music alignment aims at providing a way to navigate among these representations in a unified manner, lending itself applicable to a myriad of domains like music education, performance, enhanced listening, automatic accompaniment and so on. This project (QMUL3) has a 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 Music Information Retrieval (MIR) research through this project since work in the field of MIR has been largely systems-focused despite recurring calls for a greater focus on user-centric research.

Contribution: The extraction of music information from audio has been studied to a considerable extent in recent work in MIR, however the consideration of the performance environment and the user perception component still remains largely unexplored [6] [7] [8] [9]. A major challenge in this direction is the gathering

[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.

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. So far, our work on specifically user-centric methods for music alignment has been limited. It should be noted that our paper at EUSIPCO 2020 [57] describes a method to learn frame similarity as a neural pre-processing step for music alignment. This neural step can be leveraged to incorporate user context by providing the model with a specific type of data, which models the user's context, during training.

Impact: We foresee a significant impact on both the scientific community and the music market in the long run. The scientific community interested in music synchronization and alignment would benefit from this research, and would hopefully develop it further after the PhD. This is due to the unprecedented amount of interest generated by automatic music processing tools, coupled with the advancement in artificial intelligence; making it possible to build systems which are capable to cut the costs of human intervention on difficult tasks like transcription and alignment. The impact of the work on user-centric alignment methods will be the enhanced ability of the alignment models to be able to adapt to user context, both in terms of acoustic settings as well as the level of the user, say, in the online learning scenario. The research conducted herein could be employed for building robust systems to aid online music lessons, particularly relevant at the present time, with the spike in remote teaching and online learning. Work specifically on alignment systems will also benefit the community in multiple ways; via applications in the entertainment domain, where an on-line score alignment could be used to drive an automatic accompaniment system; to the performance domain, where it could be used for automatic page turning to aid musicians, and synchronized visualization generation to aid listeners; to the music education setting, where students could be shown where their performance deviates from indicated score markings.

Future work: 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. 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 a tablet. 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.

3.4 QMUL4: Drum Sound Query by Vocal Percussion

In this project, we explore techniques to query drum sounds using vocal imitations as input, usually plosive consonant sounds. Query by Vocal Percussion (QVP), as studied in this project, has two independent yet complementary processes. Firstly, we can find similarities between the acoustic (timbral) space of drum sounds and that of their corresponding vocal imitations so that the user of the QVP system could select the drum sample he/she wants to use just by vocalising it. We call this process sample selection. And secondly, a certain user can show the QVP algorithm enough samples of a certain vocal sound, like the syllable "pm", and let the algorithm know that he/she wants to trigger a kick drum sound anytime it is detected. We call this process performance transcription.

We dive into these two processes with the user always in mind, so that later on musicians can select drum

[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.

samples and prototype drum patterns quickly just by using their own voice. Performance transcription algorithms are mainly trained in a user-based way so as to adapt to different users with distinct imitation styles [10]. In that way, we train one algorithm for each one of the users. For sample selection, we have studied timbral descriptors that can link drum sounds with their respective vocal imitations for several users and explored the interpersonal differences in their way of imitating them [11]. Results suggest that, although a general user-agnostic strategy is effective to some extent, a user-driven approach is very likely to significantly refine accuracies. Following this hypothesis, we plan to conduct a user-driven sample selection study to let different participants train active learning algorithms that adapt to the imitation style of each particular user in an interactive and dynamic way.

In case time allows, we intend to gather all the resulting algorithms from this piece of research and make an accurate, robust, and efficient QVP system ready for musicians to use.

3.5 QMUL5: Adversarial Attacks in Sound Event Classification

QMUL5 has limited user-driven aspects. Though, one of the properties of adversarial attacks is that if you train a model to be robust against adversarial attacks the interpretability of the model improves at the cost of model performance. Current research focuses on image recognition and they show that interpretability techniques on adversarial robust models improve the visual interpretability of features [12].
We are currently extending this research for the task of singing voice detection to observe whether adversarial robustness improves the auditory interpretability of features. A model that is more interpretable can help researchers debug models better and provide more accountability on the predictions that the model makes. This, in turn, improves the user experience and user acceptance of models.

3.6 JKU1: Large-Scale Multi-modal Music Search & Retrieval without Symbolic Representations

There are several ways and forms in which music can be represented, including audio recordings, video clips, images of sheet music, and symbolic standards such as MIDI and MusicXML. Also, there has been a considerable growth of large multi-modal collections of music. Making such digital archives searchable and explorable in an intuitive, content-based way, requires the development of efficient techniques for multi-modal cross-linking (between items either from different modalities or from the same modality) and also for music identification, which is the task of retrieving the appropriate meta-information of an item, given a query in one modality. The main challenges here are the large amount of data in such collections, the fact that there is no symbolic information regarding the music items, and that there is considerable heterogeneity within the archives, comprising for example handwritten scores, and different instrumentations and musical genres. The goal of this project is to propose methods for the automatic structuring and cross-linking of large multi-modal music collections, with focus on audio recordings and sheet music images and without the need for symbolic representation, supporting tasks such as the retrieval of one modality based on another one, alignment of multiple performances to sheet music for purposes of score-based listening and comparison, and piece identification in unknown recordings. In the following, we discuss present and planned future work from the viewpoint of user-related aspects of this project.

There has been a rapid growth of large multi-modal archives of music, provided by institutions from different contexts; these include cultural institutes such as the Eliette und Herbert von Karajan Institute[1], digital libraries such as the Probado music repository [13] from the Bavarian State Library[2], concert halls such as the Vienna State Opera[3], and music publishers such as Universal Edition[4]. It could be valuable for them to make such

[1] https://www.karajan-institut.org/
[2] https://www.bsb-muenchen.de/en/
[3] https://www.wiener-staatsoper.at/en/
[4] https://www.universaledition.com/

[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.

collections searchable and explorable in a convenient way for the user, supporting tasks such as content-based analysis, synchronisation, indexing and navigation [14] of items from different modalities.
We identified in our project no direct research issues related to user interactivity. Nevertheless outcomes from our work will be highly relevant for users, and we plan to keep potential interactive scenarios in mind in case of opportunities of public demonstrations towards the end of the project.

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

Although Project JKU2 does not focus on the user perspective but rather on musical knowledge and data useful to the task, we aim at interacting with the user in proposing a real-time audio-to-score alignment that can provide additional information during a live opera performance, such as highlighting the current played bar, displaying subtitles, or providing historical context. Also, the tracker can be used as an efficient teaching tool, allowing comparing and analysing different performances of a same opera.

In Score Following, the user is directly placed at the heart of the research. This project will end with the demo of the developed tracker working in real conditions and presenting several useful information to the user. It will highlight the current bar of the score played by the artists or display the subtitle of the current sentence spoken by an actor. User feedback will be useful to design an attractive and pleasant interface.

The opera tracking system has also a great potential in terms of teaching for musicians and actors. Inspired by [15], an offline graphical user interface providing alignment between different performances of a same opera allows comparing playing style, tempo variations, voice effects, actors' acting, or conductor's movements. In this tool, we will gather all possible modalities including PDF scores, aligned audio recordings, as well as conductor's and actors' video recordings, and corresponding subtitles.

3.8 TPT1: Context Auto-Tagging for Music: Exploiting Content, Context and User Preferences for Music Recommendation

Listening to music is a common activity for many individuals on a daily basis. In many cases, we listen to different types of music based on our current activity, location, or time [16]. Previous studies showed that our context has a big influence on the music we choose to listen to [16] [17]. With the current popularity of music streaming services, many users rely on the service to play their preferred music and to discover new music that they might like. Hence, recommendation systems have gained a lot of focus from the research community for their importance in giving a better user experience [18]. Previously, the main focus of music recommendation systems was finding music that would suit the user's preferences regardless of the user's different contexts. Despite the constantly improving performance of these recommendation systems, recommending suitable music in the wrong context should not be considered a good recommendation.

Hence, there has been an increasing interest in context-aware recommendation systems recently [19]. Context-aware recommendation systems, along with classical recommendation systems, are common in many services other than music, e.g. in online shopping or movie streaming [20]. However, it is specifically more important in the case of music streaming due to the dynamic nature of listening to music. Music tracks often have duration of few minutes while users would listen to music for hours during the day. This leads to a constant need for providing recommendations to user. Additionally, the user context, e.g. activity or location, could change frequently while listening to music, which leads to changes in the user's preference and consequently needs different types of recommendation. Hence, an understanding of the different types of user contexts and their effect on the music style and listening preferences is important for improving the current state of recommendation systems.
Since accessing the user context is often not feasible due to privacy issues, allowing users to select a specific context and get recommended related tracks could be an alternative. However, for this, we need to grasp to

[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.

what extent it is possible to infer user context from audio content only. Few studies have already addressed the annotation of music datasets with user context tags [21]. However, there has been no standard procedure on how to find context tags relevant to music and how to employ them. Previous studies have focused on a number of contexts that were defined arbitrarily by the authors [16] [21].

Additionally, even scarcer research has investigated the relationship between audio content and user contexts, and the feasibility to automatically predict context from a music track's audio content [21]. Such study is important for automatically generating context-aware playlists [22] or for facilitating music discovery by context tags. Finally, studying the relationship between audio content and contexts is essential in understanding the influence of different contexts on user preference. However, this relationship highly depends on the users themselves. Hence, it is also essential to approach this problem in a user-aware setting, i.e. to study how to predict the contextual use of a track for each user independently.

Certain tags largely depend on users and their listening preferences, in particular, the tags referring to the context of music listening such as 'running' or 'relaxing' [23]. Thus, traditional auto-tagging models that rely only on the audio content without considering the case where tags depend on users, are not ideal for describing music with user-dependent tags like contexts, a challenge that we address in our proposed approach. We propose to build a user-aware auto-tagging system. Given that contextual tags are interpreted differently by different users, we hypothesize that considering the user information in training a personalized user-aware contextual auto-tagging model may help. For this, we propose to add to the system, along with the audio input, also a user input (see our publications [24], [25], [26] for more details on this work). To our knowledge, this is the first work to approach the auto-tagging problem from a user-aware perspective. Our current results already show the effectiveness of including the user in the prediction process.

3.9 TPT2: Text-Informed Lead Vocal Extraction

Singing voice separation is the task of isolating the vocals from the instrumental accompaniment in music recordings. It has user-oriented applications such as karaoke, remixing, up-mixing, and also serves as a pre-processing step for music information retrieval tasks such as singer identification or lyrics transcription and alignment. State-of-the-art performance is achieved by deep learning models trained in a supervised way which requires a data set of music mixtures along with their corresponding isolated vocal tracks. Usually, only a small amount of such data is available and the question arises how performance can be further improved without access to more audio data. This project explores the usage of lyrics as additional information for singing voice separation.

While text transcripts are often widely available for popular music, they are usually not aligned with the audio signals. Therefore, as a first step, a deep learning voice separation model is developed which can align and exploit side information jointly. As proof of concept, it is evaluated with artificial voice activity information on a singing voice separation task. Next, we perform text-informed speech music separation and text-to-speech alignment jointly. Experimental results show that this leads to mutual benefits for both tasks. Aligned text improves the perceptual speech quality and the separation objective enables alignment on highly corrupted speech. Improvements on the model are proposed to adapt it to the more challenging tasks of text-informed singing voice separation and lyrics alignment on polyphonic music. Combining attention and dynamic time warping the model aligns phonemes accurately on solo and mixed singing voice – given accurate transcripts. It also achieves competitive performance on word level alignment test sets with less accurate transcripts while being trained on much less data than state-of-the-art methods. When lyrics are aligned first and then used to inform a separation model, the separation quality can be improved.

While being mainly knowledge-driven and data-driven, the topic of this project also has some user-driven aspects. One of them is the evaluation of the separation quality. This is not the research focus of 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.

but it is nevertheless an essential part of audio source separation research. There are several ways to define “good” separation quality, which are reflected in different objective evaluation metrics such as signal-to-distortion ratio, signal-to-interference ratio, and signal-to-artefacts ratio. They are, however, not strongly related to human perception, even though this is an important aspect for users of source separation systems, unless the separation is only a pre-processing step for another task.

So far, the quality of our separation method has been evaluated with the classic objective metrics for the sake of comparability with related work, but we have also introduced novel metrics which highlight important aspect of separation quality perception by humans [27]. We also conducted small, informal listening tests asking colleagues about their opinion of the separation quality. Of course, we also listen frequently to the separated signals ourselves to gain a better understanding of their quality.

Another user-driven aspect of this work is that a text transcript has to be provided by the user in order to inform the separation. This can be a possibility to increase the control of the user over the separation beyond informing the separation system. In our most recent work, which is still in progress, we align lyrics and the audio mixture at phoneme level as a first step. Then, the voice separation is performed using the phoneme information as indicator for some expected spectral properties of the voice signal such as harmonic or noisy. We are able to modify the separated singing voice signal to a limited extent by changing some phonemes in the transcript [28] [29].

In future work this aspect could be strengthened by using a synthesis-based approach to singing voice separation. The voice signal would be synthesized from the audio mixture and text. If the singer specific characteristics and melody can be extracted from the mixture and the pronounced words generated from the text input, this would give a user control about the lyrics sung in the extracted voice signal. Such a system could be used to adapt the voice and/or content to different contexts. On the other hand, some ethical aspects would need to be considered as well: Is it desirable that someone can alter the words a singer sung in a recording? If such system works well, it needs to be assured that it will not be misused.

3.10 TPT3: Muti-modal Music Recording Remastering

The focus of this PhD is developing new methods for offering the user an improved and interactive multimedia experience while watching a movie or a video. In particular, the aim is to study methods for a user-centered remastering of music performance recordings. The idea is to guide and inform audio source enhancement algorithms using the user’s selective attention as a high-level control to select which is the desired source to extract and provide priors about it. In the case of live-music videos, the source to enhance is represented by a voice in the ensemble. Such methods could be used in the future to develop useful applications and software both for a general audience and expert users as sound engineers, video designers and musicians.

First of all, we focused on the problem of EEG-based decoding of auditory attention to a target instrument in realistic polyphonic music. We obtain promising results, showing that the EEG tracks musically-relevant features which are highly correlated with the time-frequency representation of the attended source and only weakly correlated with the unattended one. This “contrast” is then used to inform a source enhancement algorithm. The results are promising and show that EEG information is able to select automatically the desired source to enhance and improves the separation quality.

We have seen that the separation can be helped by some prior knowledge we already have. This knowledge can derive from other modalities than the audio (e.g. music score, the lyrics, the motion of the sound sources and many others). However, the listener himself is rarely considered as a potential source of information. Consider the case of a person who is listening to a piece of music and focusing on a particular instrument in the mixture. Our brain performs this task efficiently and effectively and the cognitive mechanism that makes

[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 possible is called auditory attention. This mechanism can be tracked in the neural activity and in particular using the electroencephalographic (EEG) signal. The authors of [30] [31] showed that it is possible to decode which was the attended source and possibly reconstruct an audio representation of it.

Only a few works have been proposed in the last years that combine source separation with auditory attention decoding characterized by EEG recordings [32] [33]. However, the core idea is to separate each sound source and use them as isolated sources to decode the attended speaker or, vice-versa, to decode from the EEG which source needs to be enhanced. Thus, they do not use any information from the EEG to improve the separation. [34] [35] [36] contributed to combine auditory attention decoding with beamforming techniques for speech enhancement.

However, all these works focus on speech stimuli and not on music. Taking advantage of the promising results obtained on auditory attention decoding applied to music [37], we developed a new source enhancement paradigm which can be referred to as neuro-steered music source enhancement. In [37], we have seen that the EEG tracks musically-relevant features which are highly correlated with the time-frequency representation of the attended source and only weakly correlated with the unattended one. This "contrast" represents some user-driven knowledge we can use to inform a source enhancement algorithm.

3.11 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 the taste of a particular user).

In our paper [38], we are proposing a novel system for one-shot accompaniment style transfer. Unlike most existing approaches to music style transfer where the set of possible target styles is usually very limited, our one-shot approach extracts the target style from a short example provided by the user. As a result, the user is more involved in the process, not only choosing the song they wish to transform, but also having fine control over the style of the result. To showcase the system, we developed an interactive web demo which makes it easy for users to provide inputs of their choice and adjust them to obtain the desired result.

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

Synthetic creation of sounds is commonly performed using analog or digital synthesis, allowing a musician to sculpt the desired timbre modifying various parameters. Typically, such parameters control low-level features of the sound and often have no musical meaning or perceptual correspondence. With the rise of Deep Learning, data-driven processing of audio emerges as an alternative to traditional signal processing. This new paradigm allows controlling the synthesis process through learned high-level features or by conditioning a model on musically relevant information. The goal of these projects is to study Deep Learning approaches for audio synthesis, by exploiting different sources of conditional information to gain musical control over the synthesis process. The experiments carried out in TPT5 use a Generative Adversarial Network (GAN) for the task of audio synthesis of pitched harmonic and percussive sounds. By conditioning the models on pitch or perceptual features computed with a publicly available feature-extractor, intuitive control is gained over the generation process. We also compare different audio signal representations, including the raw audio waveform and a variety of time-frequency representations. The experiments are carried out on the NSynth dataset in the case of pitched sounds, and on a large collection of kick, snare, and cymbal sounds collected from Sony CSL's private datasets. We quantitatively evaluate the generated material utilizing standard metrics for assessing generative models (Inception Score, Kernel Inception Distance, etc.), and compare training and

[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.

sampling times. We show that complex-valued as well as the magnitude and instantaneous frequency of the short-time Fourier transform achieve the best results, and yield fast generation and inversion times on the task of pitched sound synthesis. In the case of drum sound synthesis, compared to a specific prior work, our approach considerably improves the quality of the generated drum samples, and the conditional input indeed shapes the perceptual characteristics of the sounds.

In terms of user-driven methodologies, 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 [39] or, given a simple painted sketch of a landscape, render a high-quality real image version of it [40]. This project aims at applying representation learning methodologies to audio. The goal is to learn musically motivated and interactive methods for controlling audio synthesis in a musically meaningful way, avoiding complicated controls existent in modern synthesizers [41].

Learning such semantically meaningful representations is possible thanks to data-driven methods to process audio (e.g. Using Deep Learning, DL) instead of traditional signal processing methodologies (e.g. additive synthesis). This new paradigm allows us to steer the synthesis process by manipulating learned higher-level latent variables, which provide a more intuitive control compared to conventional synthesizers. In addition, as DL models can be trained on arbitrary data, comprehensive control over the generation process can be enabled without limiting the sound characteristic to that of a particular synthesis process (e.g. additive/subtractive synthesis, wavetable, etc). For example, GANs allow to control the synthesis through their latent input noise [42] and Variational Autoencoders (VAE) can be used to create variations of existing sounds by manipulating their position in a learned timbral space [43]. However, an essential issue when learning latent spaces in an unsupervised manner is the missing interpretability of the learned latent dimensions. This can be a disadvantage in music applications, where comprehensible interaction lies at the core of the creative process. Therefore, it is desirable to develop a system which offers expressive and musically meaningful control over its generated output. A way to achieve this, provided that suitable annotations are available, is to feed higher-level conditioning information to the model. The user can then manipulate this conditioning information in the generation process. Along this line, some works on sound synthesis have incorporated pitch-conditioning [44] [45], or categorical semantic tags [46], capturing rather abstract sound characteristics.

In the case of drum pattern generation, there are neural-network approaches that can create full drum tracks conditioned on existing musical material [47]. In a recent study, a U-Net is applied to neural drum sound synthesis, using conditional perceptual features as a means for controlling the timbre of the generated sound [48]. However, in that work, the neural network learns a deterministic mapping of the conditional input information to the synthesized audio. This limits the model's capacity to capture the variance in the data, resulting in a sound quality which does not seem acceptable in a professional music production scenario.

Our second paper builds upon that former idea and presents DrumGAN, a generative adversarial network for the synthesis of drum sounds offering high-level control over the timbral characteristics of the generated sounds [49]. This is achieved by conditioning the network on continuous features describing the perceived timbral characteristics (e.g., boominess, brightness, depth, etc.), computed with the Audio Commons timbre models[5].

We conduct our experiments on a dataset of a large variety of kicks, snares, and cymbals with approximately 300k samples. Our architecture is based on the Progressive Growing Wasserstein GAN [12], which has yielded

[5] https://github.com/AudioCommons/ac-audio-extractor

[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 results in adversarial audio synthesis. Also, we investigate whether the feature conditioning improves the quality and coherence of the generated audio. For that, we perform extensive experimental evaluation of our model both in conditional and unconditional settings. We evaluate our models by comparing the Inception Score (IS), the Frechet Audio Distance (FAD), and the Kernel Inception Distance (KID).

Furthermore, we evaluate the perceptual feature conditioning by checking if changing the value of a specific input feature yields the expected change of the corresponding feature in the generated output [49]. Audio samples of DrumGAN can be found on the accompanying website[6].

3.13 UPF1: Facilitating Interactive Music Exploration

Music recommendation systems are an integral part of modern music streaming services. Usually, they utilize the exploit-vs-explore model from game theory. There is a lot of work being done on music exploitation, but not enough on exploration. In project UPF1, we address the issue of strict categorization of music according to classes (genres, moods, themes) by utilizing the latent space that is learned by deep music auto-tagging systems. We introduce the web interface to directly explore latent tag and embedding spaces. We conduct user experiments to understand the semantics of the dimensions of the latent spaces. We also introduce the assistant system that uses reinforcement learning to learn the user's context and goal as well as personalize the process to provide the most relevant exploration directions at the time of interaction. The evaluation of the system is performed based on the experiments involving user interactions with the system with the metrics based on user engagement and novelty.

While this research is heavily data-driven, the application is user-centric and will be evaluated accordingly. We plan to conduct user experiments and observe how users interact with the interface to evaluate it and improve it for the purpose of the exploration.

So far we have introduced an online web interface that allows for direct exploration of the latent spaces, so it can be used for music exploration by users (see Figure below).

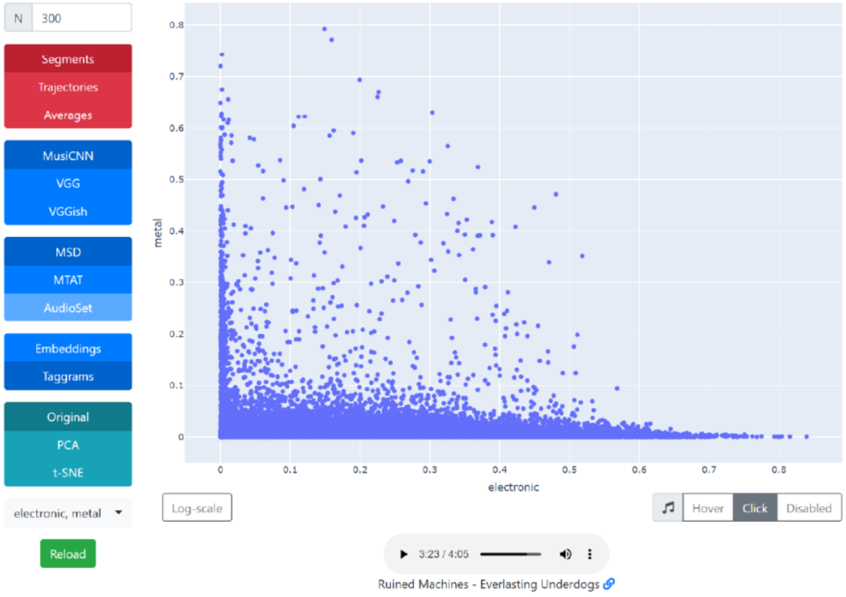

Figure 3: Exploring electronic metal

[6] https://sites.google.com/view/drumgan

[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 will also serve as one of the candidate systems in addition to the system that will purely focus on the exploration with a simplified interface with prompts that will guide the user's navigation in the latent space. The baseline will consist of the clones of the systems that are currently used in the industry: a one-button-fits-all system that will just mimic Spotify's “discover weekly” functionality, and the full catalog of genre and mood playlists that need to be browsed by the user.

The design of the proposed exploration system is very user-centric, as we plan to apply the reinforcement learning approach with the reward function determined by user engagement and feedback. We will investigate reinforcement learning approaches that work well with fast learning from the user's feedback. One optional research question that we want to address is to see if reinforcement learning systems will be able to learn the global patterns of the user in the relation to the exploration process and separate it from local variances in user mood and goals.

Moreover, for evaluation of the performance of our system, we use user engagement metrics as opposed to accuracy metrics that are typically used in recommendation system research. Qualitative evaluation is more important than quantitative in the context of assessing exploration systems, as we want users to feel that they are exploring new things, even if they do not like the recommendations from the system. We plan to use both explicit and implicit feedback from users to evaluate the exploration system.

Some of the user-centric evaluation will be done in part by collaboration with the industrial partner of this project, Jamendo. The plans include adapting the prototype to the Jamendo use-case and deploying it in an A/B testing environment that provides a great opportunity to gather a significant amount of user evaluations and feedback and rapidly iterate to deploy and test different improvements of the system.

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

The work we conducted in UPF2 which relates to user-driven methodologies is our research in the evaluation of harmonic compatibility metrics for the retrieval of audio loops. While several algorithms have been proposed for analysing the consonance and harmonicity of audio [50], these have never been evaluated for the retrieval of loops. The evaluation data consists of loops from Looperman with only harmonic content and with tempo 140 BPM. Loops from different instruments were selected and, for each of the proposed algorithms, we retrieved the 5 loops which are considered more consonant with the query loops. We firstly analysed the audio characteristics of the retrieved loops for each algorithm. Then, we evaluated user preference through an A/B test. Each of the reference sounds was presented playing together with the retrieved sounds for 2 algorithms and we asked the user which combination was preferred.

In order to make sure that the generative models we propose [51] provide the best quality for the users, we conducted listening tests with them. In both our proposed generative models, we use different loss functions, some of which use perceptually relevant losses. For each model, we conduct a listening test to evaluate which of the sounds generated by the model with different loss functions has the highest quality, in comparison to a reference. For our work with one-shot sound generation, we conduct an A/B listening test. In our work on loop generation, as we have more loss functions available, we conduct a MUSHRA-like listening test to also evaluate synthesis quality.

We aim to extend our user-driven work by analysing how the three different proposed approaches (classification, generation and compatibility) enhance creativity and workflow in music creation scenarios. For this, we will implement these algorithms in a prototype and conduct user tests and interviews with artists using the prototype.

[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.

3.15 UPF3: Encoding the Essence of Musical Compositions with Computational Approaches

Version identification (VI) is referred to as the task of detecting and retrieving a set of songs that are derived from the same underlying musical composition. Also called "cover songs", versions can be defined as any reinterpretation of existing musical works (e.g., live performances, demos, instrumental versions), and they express the same musical entity while incorporating changes in a number of musical characteristics that can be categorized in 8 main groups: timbre, key, tempo, timing, structure, harmonization, noise, and lyrics. Due to the goal of looking past such changes and somehow identifying the compositional essence of songs, VI poses a more challenging setting compared with other content-based music retrieval tasks like audio fingerprinting or near-exact duplicate detection.

Our main motivation for this project is to build VI systems that can be used in real-world scenarios. To achieve this, such systems have to be both accurate (i.e., providing reliable results) and scalable (i.e., providing fast retrieval from databases of millions of songs). While substantial steps have been taken by other VI researchers toward this goal, we plan to further improve the current state of VI systems and explore the evaluation contexts that are outside of the academic datasets.

Historically, the version identification (VI) task is addressed using knowledge- and data-driven methodologies and the role of the end-users has been mostly neglected. This can be explained with two observations:

1. the links that connect versions can be defined objectively without any user input, and
2. the successful VI systems were struggling in terms of scalability, which affects the possibility of implementing applications that would serve the end-users.

Not considering the user aspect has limited the VI research to academic and business-to-business ecosystems. Although this is a natural consequence of the fact that the main application of VI systems is digital rights management, it also reduces the potential impact of VI research. However, by bridging the accuracy-scalability gap that affected VI systems for decades, new applications of such systems can be explored. A key advancement that would facilitate exploring new technologies for VI can be developing applications that can operate in consumer electronics (e.g., smartphones and tablets). If end-users can query an audio segment and the system can retrieve the most similar documents from a database of millions of songs, the commercial value of VI research can be taken to a new level.

The goal of building accurate VI systems that can scale up to real-world data is the main motivation for our research. Therefore, our work has explored new avenues to improve the current status of VI research, by proposing state-of-the-art systems that can be exploited in a larger set of application scenarios [52], [53], [54]. Specifically, we have addressed the computation and storage requirements of previous systems and showed that it is possible to obtain a sufficient performance while having a scalable system. With our proposed systems, we hope to expand the range of applications VI systems can serve, which is an important consideration in order to realize the potential impact of VI research.

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

4. References

[1] Cannam, C., Landone, C., & Sandler, M. (2010). Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files. In Proceedings of the 18th ACM International Conference on Multimedia.

[2] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., & Silovsky, J. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding.

[3] Demirel, E., Ahlbäck, S., & Dixon, S. (2020). Automatic lyrics transcription using dilated convolutional neural networks with self-attention. In IEEE International Joint Conference on Neural Networks (IJCNN) 2020.

[4] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in Workshop on Automatic Speech Recognition and Understanding. IEEE, 2013.

[5] A. Stolcke, "Srilm - an extensible language modeling toolkit," in Seventh International Conference on Spoken Language Processing, 2002.

[6] 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.

[7] 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.

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

[9] 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.

[10] A. Delgado, S. McDonald, N. Xu, C. Saitis and M. Sandler, "Learning Models for Query by Vocal Percussion: A Comparative Study", Proceedings of the 46th International Computer Music Conference, ICMC, Santiago de Chile, Chile, 2021. (Accepted).

[11] A. Delgado, C. Saitis and M. Sandler, "Spectral and Temporal Timbral Cues of Vocal Imitations of Drum Sounds", Proceedings of the 2nd International Conference on Timbre, Thessaloniki, Greece, 2020. (Accepted).

[12] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, B. Tran, and A. Madry, "Adversarial Robustness as a Prior for Learned Representations," arXiv:1906.00945 [cs, stat], Sept. 2019.

[13] Jürgen Diet and Frank Kurth. The Probado Music Repository at the Bavarian State Library. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 501–504, Vienna, Austria, 2007.

[14] Meinard Müller, Andreas Arzt, Stefan Balke, Matthias Dorfer, and Gerhard Widmer. Cross-modal music retrieval and applications: An overview of key methodologies. IEEE Signal Processing Magazine, 36(1):52–62, 2019.

[15] Verena Thomas, Christian Fremerey, David Damm, and Michael Clausen. Slave: A Score-Lyrics-Audio-Video-Explorer. In Proc. of the International Society for Music Information Retrieval Conference (ISMIR), pages 717–722, Kobe, Japan, 2009.

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

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

[18] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. International Journal of Multimedia Information Retrieval, 7(2):95–116, 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.

[19] Norha M Villegas, Cristian Sanchez, Javier Dıaz-Cely, and Gabriel Tamura. Characterizing context-aware recommender systems: A systematic literature review. Knowledge-Based Systems, 140:173–200, 2018.

[20] V Subramaniyaswamy, R Logesh, M Chandrashekhar, Anirudh Challa, and V Vijayakumar. A personalised movie recommendation system based on collaborative filtering. International Journal of High Performance Computing and Networking, 10(1-2):54–63, 2017.

[21] Xinxi Wang, David Rosenblum, and Ye Wang. Context-aware mobile music recommendation for daily activities. In Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012.

[22] Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. Recsys challenge 2018: Automatic music playlist continuation. In Proceedings of the 12th ACM Conference on Recommender Systems, 2018.

[23] Minz Won, Sanghyuk Chun, and Xavier Serra. Toward interpretable music tagging with self-attention. arXiv preprint arXiv:1906.04972, 2019.

[24] Karim Ibrahim, Elena Epure, Geoffroy Peeters, Gael Richard "Should we consider the users in contex-tual music auto-tagging models?" in Proc. of the International Society for Music Information Retrieval (ISMIR), Oct. 2020 (accepted).

[25] Karim Ibrahim, Elena Epure, Geoffroy Peeters, Gael Richard "Confidence-based Weighted Loss for Multi-label Classification with Missing Labels." The 2020 International Conference on Multimedia Retrieval (ICMR '20), Jun 2020, Dublin, Ireland.

[26] K.M. Ibrahim, J. Royo-Letelier, E. Epure, G. Peeters, G. Richard, "Audio-Based Auto-Tagging with contextual tags for music" in Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, Barcelona, Spain.

[27] K. Schulze-Forster, C. Doire, G. Richard, R. Badeau, "Weakly Informed Audio Source Separation", in Proc. of WASPAA, New Paltz, NY, USA, 2019.

[28] K. Schulze-Forster, C. Doire, G. Richard, . Badeau, "Joint phoneme alignment and text-informed speech separation on highly corrupted speech" in Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, Barcelona, Spain.

[29] K. Schulze-Forster, C. Doire, G. Richard, R. Badeau, Phoneme Level Lyrics Alignment and Text-Informed Singing Voice Separation, submitted to IEEE Trans. on ASLP, 2020.

[30] Nima Mesgarani and Edward F Chang. Selective cortical representation of attended speaker in multi-talker speech perception. Nature, 485 (7397): 233 , 2012.

[31] James A O'sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cerebral Cortex 25(7):1697-1706, 2015.

[32] Simon Van Eyndhoven, Tom Francart, and Alexander 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.

[33] Neetha Das, Simon Van Eyndhoven, Tom Francart, and Alexander 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 1660ff.

[34] Ali Aroudi and Simon Doclo. Cognitive-driven binaural lcmv beamformer using eeg-based auditory attention decoding. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 406 – 410 . IEEE, 2019.

[35] Ali Aroudi, Bojana Mirkovic, Maarten De Vos, and Simon Doclo. Auditory attention decoding with eeg recordings using noisy acoustic reference signals. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 694ff

[36] Ali Aroudi and Simon Doclo. Cognitive-driven binaural beamforming using EEG-based auditory attention decoding. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28 : 862 – 875 , 2020.

[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.

[37] Giorgia Cantisani, Slim Essid, and Gaël Richard. EEG-based decoding of auditory attention to a target instrument in polyphonic music. In 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pages 80 – 84 . IEEE, 2019.

[38] Ondřej Cífka, Umut Şimşekli, Gaël Richard. "Groove2Groove: One-shot music style transfer with supervision from synthetic data." In review for IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP).

[39] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: CoRR abs/1812.04948 (2018). arXiv: 1812.04948. url : http://arxiv.org/abs/1812.04948.

[40] Taesung Park et al. "Semantic Image Synthesis With Spatially-Adaptive Normalization". In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 2337–2346, 2019. doi : 10.1109/CVPR.2019.00244.

[41] J. Nistal, S. Lattner, and G. Richard, "Comparing representations for audio synthesis using generative adversarial networks," in Proc. of the 28th European Signal Processing Conference, EUSIPCO2020, Amsterdam,NL, Jan. 2021.

[42] Chris Donahue, Julian McAuley, and Miller Puckette. "Adversarial Audio Synthesis". In: Proc. of the International Conference on Learning Representations, ICLR. 2019.

[43] Cyran Aouameur, Philippe Esling, and Gaetan Hadjeres. "Neural Drum Machine: An Interactive System for Real-time Synthesis of Drum Sounds". In: CoRR abs/1907.02637 (2019). arXiv: 1907.02637. url : http://arxiv.org/abs/1907.02637.

[44] Jesse Engel et al. "GANSynth: Adversarial Neural Audio Synthesis". In: Proc. of the International Conference on Learning Representations, ICLR. 2019. arXiv: 1902.08710.

[45] Jesse Engel et al. "GANSynth: Adversarial Neural Audio Synthesis". In: Proc. of the International Conference on Learning Representations, ICLR. 2019. arXiv: 1902.08710.

[46] Philippe Esling et al. "Universal audio synthesizer control with normalizing flows". In: Applied Sciences (2019).

[47] Stefan Lattner and Maarten Grachten. "High-Level Control of Drum Track Generation Using Learned Patterns of Rhythmic Interaction". In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA. New Paltz, NY, USA, Oct. 2019, p. 35

[48] A. Ramires et al. "Neural Percussive Synthesis Parameterised by High-Level Timbral Features". In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020, pp. 786–790.

[49] J. Nistal, S. Lattner and G. Richard, "DrumGAN: Synthesis of Drum Sounds With Timbral Feature Conditioning Using Generative Adversarial Networks" in Proc. of the International Society for Music Information Retrieval (ISMIR), Oct. 2020 (accepted).

[50] Harrison, P., & Pearce, M. T. (2020). Simultaneous consonance in music perception and composition. Psychological Review, 127(2), 216.

[51] Ramires, A., Chandna, P., Favory, X., Gómez, E., & Serra, X. (2020, May). Neural Percussive Synthesis Parameterised by High-Level Timbral Features. In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[52] F. Yesiler, J. Serrà, and E. Gómez. Accurate and scalable version identification using musically-motivated embeddings. In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 21–25.

[53] F. Yesiler, J. Serrà, and E. Gómez. Less is more: Faster and better music version identification with embedding distillation. In Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR), 2020 (accepted).

[54] G. Doras, F. Yesiler, J. Serrà, E. Gómez, and G.Peeters. Combining musical features for cover detection. In Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR), 2020 (accepted).

[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.

[55] C. Lordelo, E. Benetos, S. Dixon and S. Ahlbäck, "Investigating Kernel Shapes and Skip Connections for Deep Learning-Based Harmonic-Percussive Separation," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, 2019.

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

[57] Agrawal, R. and Dixon, S. (2020). Learning Frame Similarity Using Siamese Networks for Audio-to-Score Alignment. In *European Signal Processing Conference (EUSIPCO 2020)*, Amsterdam, The Netherlands.

[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.