New Frontiers in Music Information Processing (MIP-Frontiers)

Grant Agreement Number: 765068

Title:	First report on novel knowledge-driven approaches in MIR
Lead Beneficiary:	JKU
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 project proposal, the MIP-Frontiers consortium identified **three big challenges** that the Music Information Retrieval (MIR) field will have to tackle. One of these is the need for more high-level knowledge in MIR systems, and in the research and development process.

The **importance of knowledge** (in addition to *data* and *algorithms*) in MIR research and development was argued in the following paragraph in the project proposal, which also gives an (implicit, by way of examples) definition of what we mean by "knowledge" in this research context, and the different roles it can play: *"Developing systems that properly address and handle this richness requires expert knowledge from different fields, from musicology to psychology to physics and acoustics. Such additional knowledge may come from different sources and find its way into music systems in various ways, for example: knowledge about music perception encoded in advanced hand-crafted features; physical/acoustical models of instruments as a source of bias in audio separation or transcription; musicological-stylistic knowledge used as constraints on permitted solutions; or information from other modalities and information sources (video, user and performance context, etc.) that provides additional guidance and context to a music processing system. […]. Consequently, our strategy will be to combine the latest advances from (deep) machine learning with research on new ways of introducing and exploiting high-level knowledge to solve a number of complex musical tasks. Different individual research projects will do this in different ways, and in the context of different tasks."*

The present report pertains to **Work Package 2** ("Knowledge-driven Methodologies") of the project, whose general goals are laid out in the form of two tasks: *representation and models of specific kinds of of musical / acoustic background knowledge* (Task 2.1) and *methods for exploiting such knowledge to guide or constrain (deep) learning algorithms* (Task 2.2). Both of these tasks are addressed in our current work, as will be seen in the following.

The specific **purpose** of the present report is to give an **interim update on progress** made along these lines, in terms of new methods developed and approaches investigated in the context of the specific tasks that the 15 different PhD projects are tackling – which are rather different among each other and thus relate to knowledge aspects and challenges in different ways, and to different degrees. The following report is thus structured according to individual PhD projects, and largely based on the individual ESRs' own analyses of the role of background knowledge in their projects, and their contributions in this respect. As will be seen, the knowledge aspects are not equally prominent in the various projects, which will be reflected in project descriptions of varying lengths.

[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.

2. Importance of Background Knowledge and New Contributions to Knowledge-driven Methodologies (by Project)

2.1 Project QMUL1: *"Automatic Lyrics Transcription"*

Overall Goal: 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 is as yet no automatic lyrics transcription system that could reach similar robustness levels as current ASR systems. 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. To this end, we need to analyze differences in acoustic, linguistic and pronunciation between singing and speech, and develop singing-adapted models.

Knowledge-driven Aspects and Contributions: ALT is a complex problem that requires domain knowledge from several areas during its design. Pertinent fields include speech recognition, music information retrieval, natural language processing and, last but not least, machine learning. We consider the domain knowledge from these fields when designing the ALT system.

The first aspect concerns explicit *language modeling*: influenced by the prosodic elements of both natural language and music, song lyrics exhibit unique structures. Often rhyming is the priority over rules of grammar. We are establishing a specialised text corpus, to be used for creating a *lyrics language model* that captures specific knowledge about the structure of lyrics in song.

In our baseline ALT system [1], we employ neural networks, and designed our *neural network architecture* with respect to the properties and constraints that data in the singing voice domain tend to exhibit, such as language specificities and data resource availability.

In our recently submitted work [2], we have performed a computational analysis to get a better understanding of the systematic pronunciation variations in sung utterances. Using the knowledge that the study reveals, we have developed a novel *pronunciation dictionary* for singing voice. In the aforementioned study, it is shown that the singing-adapted pronunciations are beneficial for ALT.

We are currently also looking at the problem of sung lyrics in *theatrical singing*, like opera performances, from the perspective of the audio-to-audio alignment task. We hypothesize that during recitative sections the phonetic structures of the reference and target audio recordings remain similar, even though the musical content does not, due to the improvisational nature of these performances. Thus, we are working on audio-to-audio alignment in opera (recitatives) based on forced-aligned singer phonemes. This is a cooperation with project JKU2 (Charles Brazier), which works on automatic live opera tracking, where recitatives are particularly challenging passages, and where we hope to contribute to significant advances with language-informed lyrics alignment.

[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.

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

Overall Goal: Automatically transcribing polyphonic music (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 on multi-pitch estimation and instrument tracking algorithms. It is mandatory to have a way of recognising the instrument that played each note and associate each sound to the correct voice in the final staff notation. Project QMUL2 looks at exploiting 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 for both source separation and instrument recognition tasks.

Knowledge-driven Aspects and Contributions: There are a number of opportunities for using knowledge-driven approaches to enhance the performance of data-driven methods. In other words, by using domain-specific knowledge, it is possible to create more efficient and powerful machine learning algorithms.

Our first pertinent contribution was a *novel convolutional network* for performing harmonic-percussive source separation much more efficiently with greatly reduced number of parameters [1] . This was done by exploiting the fact that percussive signals form vertical patterns in the music spectrogram while harmonic signals tend to form horizontal structures. In our publication, we proposed to use filters (kernels) of vertical and horizontal shapes in the convolutional layers to make the network learn the different time-frequency patterns more efficiently. This is an example of how to develop a new data-driven approach (neural network) for source separation using a knowledge-driven approach to making a design choice.

Another instance of knowledge-driven research is the instrument recognition algorithm we are currently developing (as yet unpublished). The algorithm is based on our understanding of how both the transient and the stationary parts of a music sound influence our perception of timbre in different ways. Briefly, the transient part carries significant information about the production mode of the sound such as bow strokes, plucking or hammering of strings, breath impulsion for wind instruments, etc. Characteristics of the transient part of sounds such as attack time and shape are used by humans to discriminate different instruments. Regarding the stationary part, the relation between the energy of different harmonics and the spectral centroid have a high impact in our perception of timbre. Motivated by this insight into timbre perception, we proposed a method for frame-level instrument recognition using Transient-Stationary Source Separation as a pre-processing step. By doing so we can explicitly include the information related to the transient and the stationary parts as different inputs to the system so they can be analysed separately. We believe that this can facilitate learning of more efficient feature maps in the neural network and improve instrument recognition performance. Our initial results are promising.

For the rest of the project we will be trying to ally both types of methods – data- and knowledge-driven – in the research. The focus is to develop new deep learning architectures guided by our prior knowledge. One concrete idea will be to exploit knowledge from the field of musical acoustics (detailed models of the mechanics of instruments, and the resulting range of sounds they can produce) to help in the neural network design choice and its hyperparameter settings.

[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.

2.3 Project QMUL3: *"Leveraging User Interaction to Learn Performance Tracking"*

Overall Goal: Project QMUL3 is aimed at the development of robust music alignment models using deep learning methods. Traditional approaches to music alignment typically rely on hand-crafted features, which often fail to generalise to different instruments, acoustic environments and recording conditions. QMUL3 addresses this feature engineering bottleneck by employing deep neural networks which can capture both low-level features of relevance to the alignment task and higher-level mappings between feature sequences of corresponding performances. We work on metric learning for the alignment task and demonstrate promising results using our method. We explore data augmentation and semi-supervised learning to improve model performance.

Knowledge-driven Aspects and Contributions: QMUL3 has a moderate knowledge-driven component, combined with more dominant data-driven and user-driven components. As mentioned above, we aim at alignment models that are able to learn from data directly and are adaptable to different settings. While these methods are mainly data-driven, we employ certain domain knowledge to guide the design of our model structures. Our EUSIPCO 2020 paper [1] presents metric learning for the alignment task, where we learn frame similarities using a Siamese CNN architecture. Our method can be seen as a hybrid approach that combines knowledge-driven and data-driven components. The frame similarity computation can be understood to be a neural pre-processing step, which is then passed on to a dynamic time warping framework. In this way, we combine the advantages of both: the neural preprocessing step helps to make the model adaptable, while the DTW based step helps to capture the alignment task more optimally.

We also generate additional data samples for training via data augmentation, exploiting knowledge about the way recording techniques have changed over the decades, by using techniques such as random filters for data augmentation for training the alignment models. Further aspects of our data augmentation regime are inspired by, e.g., the knowledge that pianos are generally not perfectly in tune, which motivates specific data modification strategies.

2.4 Project QMUL4: *"Drum Sound Query by Vocal Percussion"*

Overall Goal: Project QMUL4 aims at exploring new techniques to query drum sounds from sound collections using vocal imitations (usually plosive consonant sounds) as query input. We call this process *Query by Vocal Percussion* (QVP).

Knowledge-driven Aspects and Contributions: Again, project QMUL4 has only a relatively minor knowledge-driven component. In contrast to previous work on QVP, where essentially, the systems only learn a correlation between query sounds and intended target (irrespective of the actual sound of the query, and whether this is similar to the intended instrument),we are trying to 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 sound he/she wants to use just by vocalising it [1]. We recently discovered how certain traditional timbral features can link drum sounds and their vocal imitations very well for several users, implying that these participants used these timbral cues present in drum sounds to construct the sound of their vocal imitations. Apart from that, we have explored deep learning and *data augmentation* techniques to refine performance transcription algorithms, which has not been done to date [2]. We used the recently released AVP dataset for that end [3]. In the near future, we intend to use

[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.

phoneme datasets and transfer learning techniques to map as much acoustic/phonetic knowledge to our task as possible.

2.5 Project QMUL5*: "Adversarial Attacks in Sound Event Classification"*

Overall Goal: Project QMUL5 focuses on the properties of adversarial attacks in audio applications. It started by demonstrating the presence of adversarial attacks in sound event classification, and showing that we need stronger adversarial attacks for audio. The next step is to study the transferability properties of adversarial attacks. The main focus then is on robustness issues in a singing voice detection task where there is an issue of volume sensitivity. We explore techniques to make singing voice detection systems invariant to volume sensitivity. Broadly, the goal of our work is to study the robustness of deep learning models to different types of inputs.

Knowledge-driven Aspects: QMUL5 is a mainly data-oriented project, with knowledge-related aspects mainly arising in the context of previous expertise in acoustic modeling. Deep learning models for singing voice have been shown to be sensitive to volume changes: by changing the volume the prediction of the output can be drastically changed. Previous work has used acoustic domain knowledge to address the volume sensitivity problem, e.g., by removing the 0th coefficient of the MFCC [1] or using zero-mean convolutions [2]. Our work uses this knowledge and comes up with a data-driven approach to make models insensitive to volume. The goal is to create a model that is invariant to volume by minimizing the feature distance between the original audio file and the volume changed audio file.

2.6 Project UPF1: *"Facilitating Interactive Music Exploration"*

Overall Goal: Project UPF1 is focused on researching ways to improve music exploration processes so that they can be performed in much more intuitive ways than those offered by current technology. Users should be able to find new music without scrolling through dozens and dozens of categories and playlists, and take advantage of both interactivity and personalization to improve the experience and facilitate the process itself to be enjoyable and engaging. This includes various aspects, such as exploration, discovery, interactivity. A special focus is on going beyond the common discretization of the exploration space in the form of discrete tags. according to categories. UPF1 wants to provide a way to explore a *continuous semantic space* that is still representative of all of the different music available. Moreover, we want to consider different categories of tags, such as genres, moods, and even context.

Knowledge-driven Aspects and Contributions: The knowledge-related aspects in project UPF1 relate mainly to the question of *interpretability* and *understandability* of the things learned by our models. To understand the semantics of the latent spaces learned by the data-driven approach we have introduced an online graphical web interface [2] for the exploration and evaluation of embedding and tag spaces of music auto-tagging systems. It can be useful for understanding the dimensions that are learned by auto-tagging systems and broaden the knowledge about which semantic dimensions are prevalent in the dimension-reduced spaces.

However as semantics is subjective, we also plan to conduct user experiments to understand if there is an agreement between users on the topic of the semantics of individual dimensions as well as principal components in the dimensionality-reduced space. The insights are useful 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.

explain decisions made by the exploration system to the user, so the process becomes more transparent. Moreover, for the dimensions with clearer semantics, it is also possible to give the user more explicit control over the navigation in the latent space.

An important aspect of the MTG-Jamendo dataset [1] is the separation of tags into 3 categories: genres, instruments and moods, and themes. While all categories are useful for music exploration, they describe different aspects of music. We make an effort to understand if the architectures to predict tags from different categories can be improved with the usage of the domain knowledge. The moods and themes category is of particular interest, as its tags are quite challenging (e.g. nature, fun, love, cinematic, melancholic). Thus we have organized an "Emotion and Theme Recognition in Music Using Jamendo" task in the context of the *Multimedia Evaluation Benchmark (MediaEval)* workshop. In the 2019 edition of the task, the best performances were achieved by the ensemble models consisting of different variations of the ResNet architecture [3], which usually also perform quite well on other categories, thus not providing any useful insights. The task will be repeated in the same format for MediaEval 2020, so the teams can improve on their approaches and more teams can join and try to compete to build a better auto-tagging system that works well with moods and themes.

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

Overall Goal: The goal of project UPF2 is to develop novel methods for browsing loops in large collections of sounds. To this end, it is necessary to develop methods for characterising the harmonic and rhythmic content of audio loops, to provide music makers with a way to navigate loop datasets in an intuitively meaningful way, enabling them to retrieve loops according to high-level semantic characteristics which they can readily understand.

Knowledge-related Aspects in this project are, again, mainly related to the issue of interpretability, matching our solutions to the knowledge of the targeted users. Our work on annotating the *Freesound Loop Dataset* [1] allows a straightforward retrieval of loops from Freesound, through a unified taxonomy with strong labels. This research uses knowledge obtained from the navigation in commercial collections of audio for enabling easier browsing in Freesound. Instead of doing textual searches to retrieve loops, which might have incorrect tags or be lacking sufficient annotations, users will be able to retrieve already annotated loops through metadata that is commonly used in commercial datasets – genre, key, tempo and instrumentation. The instrumentation search will be extended to all the loops in Freesound by developing a classifier which can perform this task automatically.

The work we conducted on the generation of drum sounds [2] offers two main benefits compared to existing work, the efficiency in sound generation and user-understandable control. Prior work in the generation of instrumental sounds relies on the use of deep learning models which take a long time to generate audio [3]. By using the Wave-U-Net [4], which uses a feed-forward approach for generation, we can create new drum sounds in almost real-time. Furthermore, prior work on drum synthesis does not allow for navigation through a learned latent space in an intuitive manner. In our work, we have selected timbral features which are semantically meaningful for music-makers, the envelope of the sound and the AudioCommons features [5]. An envelope generator is present in the majority of synthesisers, and music-makers are familiar with its parameters. For

[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.

developing the AudioCommons descriptors, regression models were developed by mapping user-collected ratings to timbral characteristics, which quantify semantic attributes. These are hardness, depth, brightness, roughness, boominess, warmth and sharpness. For the work in the generation of loops, we also use the onset detection functions of the kick, snare and hi-hats as an input to the model. However, this representation is easily mappable to a MIDI signal, which music-makers use.

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

Overall Goal: Version identification (VI) (or "cover song detection") is defined as the task of detecting and retrieving a set of songs that are derived from the same underlying musical composition. The overall goal of project UPF3 is to develop VI systems that can be used in real-world scenarios. To achieve this, such systems have to be both accurate (e.g., providing reliable results) and scalable (e.g., fast retrieval from databases of millions of songs). While substantial steps have recently 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 common academic datasets.

Knowledge-driven Aspects and Contributions: The first generation of version identification (VI) systems was designed purely from a knowledge-driven perspective [1, 2, 3]. To handle changes in various musical characteristics including key, tempo, structure, and timbre, VI researchers proposed hand-crafted solutions that led to attested accuracies in different evaluation contexts [2, 3]. However, the biggest setback was the computationally intensive nature of their hand-crafted components [3]. Based on the success of data-driven approaches in recent years, our methodology for building a novel VI system focuses on learning efficient representations using available data. However, while building a data-driven system, making certain design decisions based on domain knowledge can improve the performance of the resulting system. Here, we explain our work on understanding the relations among versions using hand-crafted, knowledge-driven techniques, and a number of design decisions we have incorporated based on domain knowledge for developing our data-driven VI systems.

One of the most important challenges of VI is to handle changes in musical dimensions that occur in various performances that originate from a certain composition. Previous research has qualitatively categorized these changes in 8 main groups: key, tempo, timing, structure, timbre, noise, harmonization, and lyrics [4, 5]. Following this, all knowledge-driven VI systems incorporate a number of hand-crafted operations to achieve invariance against those characteristics. However, no large-scale studies have been done to quantify the frequency and the extent of these changes. Our first contribution to knowledge-driven VI was to study and develop methods to address this need. For this, we included a subset for "cover analysis" in Da-TACOS, our publicly available dataset for VI research [6]. Using the pre-extracted features for 5,000 version pairs in this subset, we have studied the changes in key, tempo, timing, structure, and semantics. Specifically, the goal was to understand how many of these version pairs are different with respect to these characteristics, and how drastic those differences are. While investigating differences in key and tempo is quite straightforward, we proposed novel methods to analyze changes in timing, structure, and semantics. With this work, we aimed to attract some attention to the understanding aspect of version identification 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.

To develop a novel VI system, we have followed a data-driven approach using the insights produced by years of knowledge-driven VI research. For selecting an input representation, we have decided to use tonal features (i.e., crema-PCP), which was a decision based on the empirical success of systems that exploit tonal information. In later studies, we have shown that the systems that use tonal features outperform the ones that use melodic features, even for data-driven approaches [7].

Apart from the input representation, our decisions on building the network architectures and training strategies were based on the common challenges of VI [8]. Instead of following a task-agnostic approach, we have incorporated techniques that would specifically address the variations observed among versions. Four main points we considered were pitch transpositions, micro-timing variations, and changes in tempo and structure. For developing representations that are invariant to each of these points, we used the previous literature in VI to develop strategies against them. Our ablation studies proved that each of these individual considerations improved the performance of our models. These results show that although data-driven methodologies yield successful systems, task-specific approaches using domain knowledge can carry those systems even further.

2.9 Project TPT1: *"Context Auto-Tagging for Music: Exploiting Content, Context and User Preferences for Music Recommendation"*

Overall Goal: Project TPT1 aims at improving music recommendation systems by integrating contextual information about a user in the recommendation process. A user's contextual information is defined as the external factors – such as activity, location, time of day, social context, etc. – that affect the user's music preferences at any given time. The aim of the project is to rely on the raw audio data plus the contextual information to provide recommendations that would suit the user's taste and current context.

Knowledge-driven Aspects: The main knowledge-related aspect in this project (up to now) is our attempt at exploiting knowledge from previous studies to establish a new research dataset that reflects potentially relevant dimensions in a proper way. There have been many studies on the relationship between music preferences and a user's context. For example, North et al. [1] studied the influence on 17 different listening situations on music preferences. A similar study [2] categorized different listening contexts into 3 categories: personal, leisure, and work. They further expanded each category into subcategories that are more specific to the situation. For example, personal is split into three subcategories: personal–being (e.g. sleeping or waking up), personal–maintenance (e.g. cooking or shopping), and personal–travelling (e.g. driving or walking). Similarly, Sloboda [3] studied the functions or purpose of listening to music for different users. However, while there have been several cognitive studies on the relationship between music and context, they have not been utilized properly in building contextual auto-tagging systems. Hence, in our approach we make use of this previous knowledge in our dataset creation phase. We thoroughly collected all the different contexts that have been studied previously in order to build an inclusive dataset that will be useful for future 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.

2.10 Project TPT2: *"Text-informed Lead Vocal Extraction"*

Overall Goal: Project TPT2 focuses on singing voice separation – 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. Project TPT2 explores the usage of lyrics as additional information for singing voice separation.

Knowledge-driven Aspects and Contributions: This project is by its very nature knowledge-oriented: it investigates how prior knowledge about the singing voice can be exploited by data-driven machine learning methods for singing voice separation. Specifically, so far, we studied voice activity information [1] and lyrics (text) [2] as prior knowledge. With the developed method, the knowledge could be exploited for the separation task even if the audio signal to be separated and the additional information were not synchronized time-wise.

Generally speaking, lyrics contain information about some aspects of the sounds produced by a voice and the order in which they appear. Text can be decomposed into phonemes – the smallest sound units of a language. Each phoneme and phoneme class have distinct acoustic and spectral characteristics that allow drawing conclusions about their properties in spectrograms and wave forms [3]. However, the connection between a phonetic transcription of lyrics and the actual singing voice audio signal is not trivial. Therefore, this connection is learned on data. Specifically, we developed a multi-modal audio source separation model that jointly aligns and exploits a sequence of side information [1]. It learns to evaluate which part of the given side information is useful for the separation of a certain audio frame and learns how to use this side information to perform the separation. In general, we observed that the integrated knowledge helped to perform the separation task, especially in difficult conditions such as little available training data [1] or low signal to noise ratios [2]. One potential impact of this research is to demonstrate alternatives to purely data-driven methods, which relaxes constraints regarding the quality and quantity of training data required.

One possible direction for future work on knowledge-driven aspects is the integration of domain knowledge regarding the *voice production process*. The physiological process to produce speech and singing signals is well understood [3]. The challenge is to model this process in a form that is suitable for data-driven methods. This step has the potential to make singing voice separation more robust because it can constrain the space of allowed voice signal estimates. To date, without such a constraint some separated voice signals sound unnatural or contain sounds that are impossible to produce by humans.

2.11 Project TPT3: *"Multimodal Music Recording Remastering"*

Overall Goal: The focus of TPT3 is on 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-centred remastering of music performance recordings. The idea is 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.

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 kind of 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.

Knowledge-related Aspects: The knowledge-related aspects in this project mainly relate to the fact that we wish to use knowledge (or hypotheses) about a user's attention to guide source separation and enhancement algorithms. First of all, we focused on the problem of EEG-based decoding of auditory attention to a target instrument in realistic polyphonic music. A special training/evaluation dataset was compiled and published for this purpose [1]. We obtained 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 [2]. This "contrast" is then used to inform a source enhancement algorithm [3]. The results are promising and show that EEG information is able to automatically select the desired source to enhance and improves the separation quality.

2.12 Project TPT4: *"Context-driven Music Transformation"*

Overall Goal: The aim of project TPT4 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).

Knowledge-driven Aspects and Contributions: Approaches proposed so far for music style transformation have mostly been fully data-driven, restricting them to be unsupervised due to a lack of aligned training data. Such unsupervised approaches are often difficult to control and may yield "unexpected" solutions which do not make sense musically.

In our two papers [1, 2], we are proposing a combination of data-driven and knowledge-driven approaches, which consists of generating a *synthetic training dataset* and subsequently using it to train a neural network. The dataset has the following properties:

1. It is generated based on human-defined rules and patterns (exploiting a commercial accompaniment generation software).
2. It is categorized into a large number (thousands) of narrowly defined style classes.
3. It is parallel, i.e. contains examples for translation from one style to another.

Therefore, it deeply captures existing knowledge about musical styles in a way which is unparalleled in existing datasets. The neural network, trained in a data-driven way, is then able to exploit this knowledge while also generalizing beyond this synthetic dataset.

In the future, the proposed approach may be extended to use a combination of synthetic and non-synthetic datasets (making it more data-driven). This should improve the generalization capabilities of the system (and hence its performance on non-synthetic inputs), while still enabling it to draw on knowledge encoded in the synthetic data.

[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.

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

Overall Goal: The goal of TPT5 is to study Deep Learning (DL) approaches for audio synthesis and, in general, music production tools, that exploit different sources of conditional information to gain musical control over the generation process. Specifically for the current stage of the project, we focus on the use of Generative Adversarial Networks (GANs) to synthesize musical audio content given prior descriptive information (e.g., pitch, instrument, and timbre features).

Knowledge-related Aspects: From a knowledge perspective, this project is about the impact of various audio representations, including the raw audio waveform and several time-frequency representations, for the task of audio synthesis with GANs. In recent years, deep learning for audio has moved away from feeding in hand-crafted features requiring prior knowledge, in favor of features learned from raw audio data or mid-level representations such as the Short-Time Fourier Transform (STFT) [1]. This has allowed us to build models requiring less prior knowledge, yet at the expense of data, computational power, and training time [9]. For example, deep autoregressive techniques working directly on raw audio [7], as well as on Mel-scaled spectrograms [8], currently yield state-of-the-art results in terms of quality. However, these models can take up to several weeks to train on a conventional GPU, and also, their generation procedure is too slow for typical production environments. On the other hand, Generative Adversarial Networks (GANs) [3] have achieved comparable audio synthesis quality and faster generation time, although they still require long training times and large-scale datasets when modeling low or mid-level feature representations [2, 5]. Therefore, in [6] we compared different audio signal representations, including the raw audio waveform and a variety of time-frequency representations, for the task of adversarial audio synthesis with GANs. To this end, we adapted several objective metrics, initially developed for the image domain, to audio synthesis evaluation. Furthermore, we investigate whether global attribute conditioning can improve the quality and coherence of the generated audio. We performed extensive experimental evaluation when conditioning our models on the pitch information, as well as in a fully unconditional setting. We used a vanilla Progressive Growing Wasserstein GAN built upon convolutional blocks [4], and found that "*complex*" and "*mag-if*" yield the best quantitative metrics, which is also aligned with informal listening of the generated samples. This is interesting, and we are not aware that *complex* was used before in audio generation. We also found that evaluation metrics are generally aligned with perceived quality, but in some cases they can be sensitive to non-audible representation-specific artifacts, or yield figures which seem over-optimistic when listening to the examples.

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

Overall Goal: The goal of project JKU1 is to develop methods for the automatic structuring and cross-linking of large multi-modal music collections, with a 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.

[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.

Knowledge-driven Aspects and Contributions: The inherent multimodal component of the project makes it strongly knowledge-based. We work mainly with audio recordings and their respective score images, and the entire process of data annotation, (partly automated) preparation and augmentation exploits musical knowledge in various ways. For instance, in automatic score page segmentation, we use knowledge about typical piano sheet music structure in order to properly segment a page into systems consisting of two staves. This results in an unrolled representation of a score which is more suitable for the machine learning pipeline. As for the audio side, we introduce a musically motivated augmentation strategy that expands our training data to different piano sounds and tempi, in an attempt to generalise our model to unseen recordings.

Early on in the project, we identified that a state of the art approach for audio-to-sheet retrieval [1] had limitations when handling tempo changes in audio recordings. Guided by the fact that global and tempo variations are frequent in music, we proposed a system that can learn robust tempo-invariant representations for audio-sheet musical retrieval [2], based on an additional soft-attention mechanism that allows for the network to decide by itself the appropriate temporal context for a given audio input by computing an attention mask. Moreover, we were able to observe and musically interpret the behaviour of the learned attention mask by analyzing changes in its entropy values.

In our current work, we attempt to retrieve corresponding musical documents within a collection, given a music piece as input, by exploiting inherent temporal relationships between consecutive embeddings. Although diverse and distinct note events can emerge from within a music signal, portions of local information from a short excerpt of an audio recording are likely to be similar [3]. In order to exploit such temporal dependencies we take inspiration from chroma-based retrieval methodologies [4] and employ an alignment algorithm based on dynamic time warping (DTW) to align sequences of embedding vectors generated by projecting queries and items from a database onto the learned embedding space.

A musical aspect which is frequent in the audio-to-score scenario is the presence of structural mismatches or alternatives, mainly caused by jumps and repetitions. A conventional alignment algorithm would fail under this scenario, so we plan to extend our system by introducing additional information regarding positions in the scores where structural changes might occur [5].

2.15 Project JKU2: “*Live Tracking and Synchronisation of Complex Musical Works via Multi-modal Analysis*”

Overall Goal: Project JKU2 aims at developing a robust and accurate audio-to-score alignment system that is able to track full, live opera performances in real time. Recent approaches that have proven their efficiency at tracking orchestral works showed multiple weaknesses when faced with operas. JKU2 wishes to develop new methods to integrate different extra-musical knowledge sources (e.g., singing voice and speech detection, acoustic event detection, video (when available)) into audio-based music score following algorithms in order to achieve robustness in tracking such large, complex, heterogeneous stage works, permitting applications in the real world of opera houses and live streaming.

Knowledge-driven Aspects and Contributions: The knowledge-related aspects in this project revolve around the idea of exploiting general knowledge about the opera genre and operatic singing, and using various additional sources of information – beyond the (live) audio signal itself

[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.

– to stabilise the tracking process and make it more flexible, in order to deal with the manifold unexpected (and sometimes extra-musical) things happening in an opera. In the work performed up to now and in ongoing and planned research, this is proceeding along the following lines:

(1) *Acoustic event detection*: Compared to orchestral performances, operas have a more complex and sometimes unpredictable structure. They are frequently interrupted by breaks, applause, interludes, acting passages, and noises of varying kinds, that principally appear in between two consecutive pieces but can also appear within the piece. The spontaneity of these events makes it impossible to predict their occurrence. To deal with this, we built audio event classifiers to detect applause, speech/singing voice, and music in the live audio stream. We demonstrated a way to combine these detectors with the On-Line Time Warping algorithm [2] to directly control the tracking process, halting the tracker during events or passages that are not thought to be part of the score. Our SMC 2020 paper [1] demonstrates that this combination is essential to gain robustness, permitting to reduce the tracking error considerably.

(2*) Acoustic knowledge about opera singing*: Compared to orchestral works, operas tend to include a dominant vocal component. In most cases, the instrumental part is reliable enough for accuracy tracking using features designed for orchestra tracking [3]. However, there remain passages – especially recitative sections – that are of a more improvisatory nature, with a musical accompaniment that does not strictly follow a score. We focused on these passages by carrying out a feature study based on the acoustic knowledge of the singing voice in opera. Considering the frequency range of the singers' voice and its singing formant phenomenon [4], we combined two trackers working in parallel, one relying on speech-sensitive features, the other on music-sensitive features, to improve the accuracy of the system [5].

(3) *Lyrics tracking and keyword spotting*: Thanks to the recent development of a large synchronized audio-lyrics dataset [6], new models have shown promising results at tracking lyrics of varied musical genres [7] or at spotting words in songs [8]. We plan to extend the tracking to opera performances in using a model pre-trained on DALI and adapted to operas thanks to new opera-specific annotations. This is part of an ongoing cooperation with project QMUL1 (Emir Demirel), which started in March 2020, in the context of the MIP-Frontiers "Sandbox Event".

(4) *Multi-modal streams:* Considering the available data for this project (from the Vienna State Opera, we received not only audio recordings, but also corresponding video recordings from various angles) and the recent advances in multi-modal studies in MIR [9], we remain open to any information retrieval from the video stream; including event detection, conductor gesture following, or instrument playing activity that can help the tracker to increase its robustness.

[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. Conclusion

This report has given an overview of the various roles that domain-specific background knowledge and information extracted from extra-musical sources can play in various MIR areas and application problems. Given the heterogeneous nature of our project, which is structured into 15 rather different and independent sub-projects (PhD theses), this overview has necessarily taken a heterogeneous form as well, displaying highly different levels of knowledge exploitation, different kinds of 'knowledge', and different approaches to using and analyzing it. As stated in the introductory section, the project-by-project accounts are based on reports we asked our PhD students (ESRs) to give on these aspects. By putting part of this reporting responsibility into their hands, we tried to make them more aware of these aspects of their work, helping them to develop a meta-view of what they are doing, and how their work fits into the general MIR research landscape and its high-level methodologies. This should also help them maintain a structured view of their ongoing and future work, and the resources they could exploit in solving their respective problems.

[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 (by project)

4.1 QMUL1

[1] Demirel, E., Ahlbäck, S. and Dixon, S. (2020). Automatic Lyrics Transcription Using Dilated Convolutional Neural Networks with Self-attention. *In IEEE International Joint Conference on Neural Networks (IJCNN 2020),* Glasgow, Scotland.

[2] Demirel, E., Ahlbäck, S. and Dixon, S. (2020). Computational Analysis and Modelling of Pronunciation in Singing Voice. Submitted to *International Conference on Computational Linguistics (COLING 2020).*

4.2 QMUL2

[1] Lordelo, C., Benetos, E., Dixon, S. and Ahlbäck, S. (2019). 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.

4.3 QMUL3

[1] 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.

4.4 QMUL4

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

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

[3] Delgado, A., McDonald, S., Xu, N. and Sandler, M. (2019). A New Dataset for Amateur Vocal Percussion Analysis. In *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound*, Nottingham, United Kingdom.

4.5 QMUL5

[1] Lehner, B., Schlüter, J. and Widmer, G. (2018). Online, Loudness-invariant Vocal Detection in Mixed Music Signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1369-1380.

[2] Schlüter, J. and Lehner, B. (2018). Zero-mean Convolutions for Level-invariant Singing Voice Detection. In *Proceedings of ISMIR 2018*, pp. 321-326, Paris, France.

4.6 UPF1

[1] Bogdanov, D., Won, M., Tovstogan, P., Porter, A. and Serra, X. (2019). The MTG-Jamendo Dataset for Automatic Music Tagging. In *Proceedings of the Machine Learning for Music Discovery Workshop*, 36th International Conference on Machine Learning (ICML 2019), Long Beach, California, USA.

[2] Tovstogan, P., Serra, X. and Bogdanov, D. (2020). Web Interface for Exploration of Latent and Tag Spaces in Music Auto-tagging. In *Machine Learning for Media Discovery Workshop, ML4MD*, Thirty-seventh International Conference on Machine Learning (ICML 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.

[3] He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA.

4.7 UPF2

[1] Ramires, A., Font, F., Bogdanov, D., Smith, J. B. L., Yang, Y.-H., Ching, J., Chen, B.-Y., Wu, Y.-K., Hsu, W.-H. and Serra, X. (2020). The Freesound Loops Dataset and Annotation Tool. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, Montreal, Canada.

[2] Ramires, A., Chandna, P., Favory, X., Gómez, E. and Serra, X. (2020). Neural Percussive Synthesis Parameterised by High-Level Timbral Features. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 786-790), Barcelona, Spain.

[3] Aouameur, C., Esling, P. and Hadjeres, G. (2019). Neural Drum Machine: An Interactive System for Real-time Synthesis of Drum Sounds. In *Proceedings of the 10th International Conference on Computational Creativity (ICCC)*.

[4] Stoller, D., Ewert, S. and Dixon, S. (2018). Wave-U-Net: A Multi-scale Neural Network for End-to-end Audio Source Separation. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France.

[5] Pearce, A., Brookes, T. and Mason, R. (2017,). Timbral Attributes for Sound Effect Library Searching. In *2017 AES International Conference on Semantic Audio*. Audio Engineering Society (AES).

4.8 UPF3

[1] Marolt, M. (2006). A Mid-level Melody-based Representation for Calculating Audio Similarity. In *Proc. of the International Conf. on Music Information Retrieval (ISMIR)*, 2006, pp. 280–285.

[2] Ellis, D. and Poliner, G. (2007). Identifying ‘Cover Songs’ with Chroma Features and Dynamic Programming Beat Tracking. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, vol. IV, pp. 1429–1432.

[3] Serrà, J., Serra, X. and Andrzejak, R. (2009). Cross Recurrence Quantification for Cover Song Identification. *New Journal of Physics*, 11:093017, 2009.

[4] Gómez, E. and Herrera, P. (2006). The Song Remains the Same: Identifying Versions of the Same Piece using Tonal Descriptors. In *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2006, pp. 180–185.

[5] Serrà, J., Gómez, E. and Herrera, P. (2010). Audio Cover Song Identification and Similarity: Background, Approaches, Evaluation, and Beyond. In *Advances in Music Information Retrieval*, Studies in Computational Intelligence, Vol. 16, chap. 14, pp. 307–332. Berlin, Germany: Springer.

[6] Yesiler, R., Tralie, C., Correya, A., Silva, D., Tovstogan, P., Gómez, E. and Serra, X. (2019). Da-TACOS: A Dataset for Cover Song Identification and Understanding. *In Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2019, pp. 327–334.

[7] Doras, G., Yesiler, R., Serrà, J., Gómez, E. and Peeters, G. (2020). Combining Musical Features for Cover Detection. In *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2020 (in print).

[8] Yesiler, F., Serrà, J. and Gómez, E. (2020). Accurate and Scalable Version Identification Using Musically-motivated Embeddings. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2020*, Barcelona, Spain, pp. 21–25.

[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.

4.9 TPT1

[1] North, A. and Hargreaves, D. (1996). Situational Influences on Reported Musical Preference. *Psychomusicology: A Journal of Research in Music Cognition*, 15(1-2):30, 1996.

[2] Sloboda, J., O'Neill, S. and Ivaldi, A. (2001). Functions of Music in Everyday Life: An Exploratory Study Using the Experience Sampling Method. *Musicae Scientiae*, 5(1):9-32, 2001.

[3] Sloboda, J. (1999). Everyday Uses of Music Listening: A Preliminary Study. *Music, Mind and Science*, pages 354-369, 1999.

4.10 TPT2

[1] Schulze-Forster, K., Doire, C., Richard, G. and Badeau, R. (2019). Weakly Informed Audio Source Separation. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 273-277. IEEE, 2019.

[2] Schulze-Forster, K., Doire, C., Richard, G. and Badeau, R. (2020). Joint Phoneme Alignment and Text-informed Speech Separation on Highly Corrupted Speech. *In 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020)*, Barcelona, Spain.

[3] Rabiner, L. and Schafer, R. (2010). *Theory and Applications of Digital Speech Processing*. Prentice Hall

4.11 TPT3

[1] Cantisani G., Trégoat G., Essid S., and Richard G. (2019). MADEEG: An EEG Dataset for Decoding Auditory Attention to a Target Instrument in Polyphonic Music. *Workshop on Speech, Music and Mind 2019 (SMM 2019)*, Vienna, Austria.

[2] Cantisani G., Essid S., and Richard G. (2019). EEG-based Decoding of Auditory Attention to a Target Instrument in Polyphonic Music. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2019),* New Paltz, NY, USA.

[3] Cantisani G., Essid S., and Richard G. (2020). *EEG-informed Source Enhancement of a Target Instrument in Polyphonic Music*. (Journal paper in preparation).

4.12 TPT4

[1] Cífka, A., Şimşekli, U. and Richard, G. (2019). Supervised Symbolic Music Style Translation Using Synthetic Data. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[2] Cífka, O., Şimşekli, U. and Richard, G. (2020). Groove2Groove: One-shot Music Style Transfer with Supervision from Synthetic Data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), 28, pp. 2638-2650*.

4.13 TPT5

[1] Dieleman, S. and Schrauwen, B. (2014). End-to-end Learning for Music Audio. In *Proceedings of ICASSP 2014*. Florence, Italy, May 2014, pp. 6964-6968.

[2] Donahue, C., McAuley, J. and Puckette, M. (2019). Adversarial Audio Synthesis. In *Proc. of the International Conference on Learning Representations, ICLR. 2019*.

[3] Goodfellow, I. et al. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014).* Montreal, Canada. pp. 2672-2680.

[4] Karras, T. et al. (2017). Progressive Growing of GANs for Improved Quality, Stability, and Variation. *ICoRR abs/1710.10196 (2017).* arXiv: 1710.10196. url: http://arxiv.org/abs/1710.10196.

[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] Marafioti, A. et al. (2019). Adversarial Generation of Time-Frequency Features with Application in Audio Synthesis. CoRR *abs/1902.04072 (2019).* arXiv:1902. url: http://arxiv.org/abs/1902.04072

[6] Nistal, J., Lattner, S. and Richard, G. (2020). Comparing Representations for Audio Synthesis Using Generative Adversarial Networks. In *European Signal Processing Conference (EUSIPCO 2020)*.

[7] van den Oord, A. et al. (2016). WaveNet: A Generative Model for Raw Audio. *CoRR abs/1609.03499 (2016)*. arXiv: 1609.03499.

[8] Vasquez, S. and Lewis, M. (2019). MelNet: A Generative Model for Audio in the Frequency Domain. *CoRR abs/1906.01083 (2019)*. arXiv: 1906.01083.

[9] Zhu, Z., Engel, J. and Hannun, A. (2016). Learning Multiscale Features Directly from Waveforms. In *Proceedings INTERSPEECH 2016*. San Francisco, CA, USA, Sept. 2016, pp. 1305-1309.

4.14 JKU1

[1] Dorfer, M., Hajic jr., J., Arzt, A., Frostel, H. and Widmer, G. (2018). Learning Audio–Sheet Music Correspondences for Cross-modal Retrieval and Piece Identification. *Transactions of the International Society for Music Information Retrieval*, 1(1).

[2] Balke, S., Dorfer, M., Carvalho, L., Arzt, A. and Widmer, G. (2019). Learning Soft-attention Models for Tempo-invariant Audio–Sheet Music Retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 216–222, Delft, Netherlands.

[3] Müller, M. (2015). *Fundamentals of Music Processing*. Springer Verlag.

[4] Balke, S., Arifi-Müller, V., Lamprecht, L. and Müller, M. (2016). Retrieving Audio Recordings Using Musical Themes. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 281–285, Shanghai, China.

[5] Fremerey, C., Müller, M. and Clausen, M. (2010). Handling Repeats and Jumps in Score-performance Synchronization. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 243–248, Utrecht, The Netherlands.

4.15 JKU2

[1] Brazier, C. and Widmer, G. (2020). Towards Reliable Real-Time Opera Tracking: Combining Alignment with Audio Event Detectors to Increase Robustness. In *Proc. of the Sound and Music Computing Conference (SMC 2020)*, pages 371–377, Turin, Italy.

[2] Dixon, S. (2005). An On-Line Time Warping Algorithm for Tracking Musical Performances. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI 2005)*, pages 1727–1728, Edinburgh, Scotland, UK.

[3] Gadermaier, T. and Widmer, G. (2019). A Study of Annotation and Alignment Accuracy for Performance Comparison in Complex Orchestral Music. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR 2019)*, pages 769–775, Delft, The Netherlands.

[4] Sundberg, J. (1977). The Acoustics of the Singing Voice. *Scientific American*, 236(3):82–91, 1977.

[5] Brazier, C. and Widmer, G. (2020). Addressing the Recitative Problem in Real-Time Opera Tracking. In *Proceedings of the 25th International Symposium on Frontiers of Research in Speech and Music (FRSM 2020)*, Silchar, India.

[6] Meseguer-Brocal, G., Cohen-Hadria, A. and Peeters, G. (2018). DALI: A Large Dataset of Synchronized Audio, LyrIcs and Notes, Automatically Created Using Teacher-student Machine Learning Paradigm. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR 2018)*, pages 431–437, Paris, France.

[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] Gupta, C., Yılmaz, E. and Li, H. (2020). Automatic Lyrics Alignment and Transcription in Polyphonic Music: Does Background Music Help? In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 496–500, Barcelona, Spain.

[8] Vaglio, A., Hennequin, R., Moussallam, M., Richard, G. and d'Alché Buc, F. (2020). Audio-Based Detection of Explicit Content in Music. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 526–530, barcelona, Spain.

[9] Duan, Z., Essid, S., Liem, C., Richard, G. and Sharma, G. (2019). Audiovisual Analysis of Music Performances: Overview of an Emerging Field. *IEEE Signal Processing Magazine* 36(1):63–73.

[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.