New Frontiers in Music Information Processing (MIP-Frontiers)

Grant Agreement Number: 765068

- Title: First report on novel data-driven approaches in MIR
- Lead Beneficiary: UPF
- 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 project had a particular emphasis on developing data-driven and system level solutions of interest to a wide variety of companies working within the audio and music field. For this goal we need large and appropriate corpora/datasets, state of the art audio/music feature analysis tools, and machine learning and evaluation strategies adequate for the specific problems identified. This report reflects that emphasis and reports on the results obtained so far.

This deliverable is part of WP1, which is focused on data-driven methods. As part of this WP there are two major tasks. Task 1.1 aims at developing efficient and scalable signal processing methods for computing high time-resolution and note-level audio features, investigating features specific to the singing voice, and evaluating on industry-provided large-scale datasets. Task 1.2 aims to investigate methods for exploiting multimodal data, including user behaviour and context, with unsupervised learning algorithms for MIR tasks; to monitor on-going progress in deep learning; to transfer and adapt promising (including crossmodal) approaches to the music processing field; and to develop network architectures for generic audio (music) processing.

This particular deliverable is a first report on novel data-driven approaches in MIR, summarizing methods and results to date from all the ESR projects.

2. Data-driven Methods in MIP-Frontiers

The first fundamental challenge, and opportunity, faced by present-day MIR is the ever-growing amount of data available with a potential to be processed and made sense of. MIP-Frontiers is working on research projects in which ESRs address music problems for which new audio signal processing and machine learning methodologies have to be developed. We take advantage of existing large data repositories, some open and others supplied by the industrial partners, and we work on the specific problems related to the complexity of the types and characteristics of the data sources.

The next section goes over all projects in MIP-Frontiers emphasizing the data-related aspects.

3. Novel data-driven approaches in MIR (by PhD Project)

3.1 *QMUL1: Automatic Lyrics Transcription*

Machine learning methods have been widely used in recent decades to solve information retrieval and pattern recognition problems. More recently, a specific branch of machine learning, *deep learning* has shown an exponentially growing interest among both researchers and commercial entities due to them outperforming other pre-existing methods. This is possible only in the presence of sufficient training data, and thus deep learning methods are fundamentally data dependent. Our research also employs deep learning methods to solve the automatic lyrics transcription problem. Therefore, data-driven approaches shape the core aspects of this project. Below is the list of data-driven aspects:

- The format of the training data: Our framework requires monophonic audio recordings with a single channel. The system processes audio at 16 kHz and 16 bit, but recordings with different sampling rates and bit rates can be reformatted. For the language model, a text corpus with sentence or lyric-line level of granularity for each training sample is needed. The text data is processed to filter non-alphabetic characters and symbols.

[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.

- The format of the annotations for training: Each training audio sample has to be at sentence or lyrics-line level of granularity, as for the text data for language models. If multiple sentences or lyric-lines are present in the training audio, their beginning and ending time-stamps are required.
- By defining the above data formats, we aim to set the standards for data pre-processing in ALT research.
- ALT research has been suffering from a lack of benchmark evaluation sets. Alongside one of our publications, we plan to release sentence-level annotations of an existing open-source dataset for singing voice, providing a new open-source evaluation set for ALT.
- Open source data sets for ALT research are limited in size compared to those for ASR. Those that have sufficient size for training are either weakly-labelled, noisy or contain polyphonic music in the background. These aspects of the available data sets require a careful design of the neural network architecture so that it can learn efficiently from small data and be robust to noise.
- In our research we consider two types of music data: the first is monophonic singing voice recordings recorded with a mobile app (ScoreCloud Express - https://scorecloud.com/, or Smule App - https://www.smule.com/), and the second type is commercial recordings with background music as accompaniment. Each type mentioned here comes with specific challenges with regard to data processing. We build our research pipelines with respect to these challenges in order to be able to demonstrate a use case for the common real-world cases in industrial applications.
- The performance of ALT systems heavily relies on music segmentation. We employ a number of supervised and unsupervised segmentation approaches depending on the data structures mentioned above.
- In our thesis, we demonstrate the application of ALT in multiple languages and ways to achieve good performance in the presence of limited resources for training data.
- In the finalized version of this project, we aim to develop a tool that could automatically generate lyrics annotations given a few minutes length of singing voice recording and corresponding lyrics. The lyrics can be retrieved from online resources or other content providers and collaborators. This system will leverage the ALT system presented in [1] and use it to curate new open source datasets for lyrics transcription.

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

The Project QMUL2 sees an opportunity of 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.

For instance, the first contribution of the project was the proposal of a novel convolutional network for performing harmonic-percussive source separation much more efficiently with greatly reduced number of parameters [2]. 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. Thus, it is possible to say that we developed a new data-driven approach (neural network) for source separation using a knowledge-driven methodology to guide our design choice.

Another combination of knowledge-driven and data-driven approaches utilised by the project is the instrument algorithm we have proposed. Even though we have not published anything regarding it yet, we have already made experiments and presented it at the Virtual MIP Frontiers workshop.

[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 better understand our approach, it is important to know that the transient and the stationary parts of music sounds influence our perception of timbre in different ways. The transient parts of sounds carry significant information about the production mode of the sound such as bow strokes, plucking or hammering of strings and breath impulsion for wind instruments. Characteristics of the transient parts of sounds such as the attack-time and shape carry essential features that are used by humans to discriminate different instruments. Regarding the stationary part, the relation between the energy of different harmonics and the spectral centroid has a high impact on our perception of timbre too.

Motivated by this previous knowledge of timbre perception, we proposed a data-driven 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 of the instrumental sounds as different inputs to the system so they can be analysed separately. We believe that this approach can facilitate learning of more efficient feature maps by the neural network and improve instrument recognition performance. Our initial results are promising.

For the rest of the PhD we will always be trying to ally both types of methods in the research. The focus is to develop novel deep-learning architectures guided by our previous knowledge. One idea we have is 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 aid neural network design choices and hyperparameter settings.

Moreover, we are planning to create a new dataset based on instrument taxonomies for instrument recognition. Our current idea is to synthesize audio from MIDI files using professionally graded synthesizers. Such a dataset would help foster research on novel data-driven methods by the MIR community.

3.3 *QMUL3: Leveraging User Interaction to Learn Performance Tracking*

The QMUL3 project focuses on music alignment. Music alignment aims at providing a way to navigate among multiple representations of music in a unified manner, lending itself applicable to a myriad of domains like music education, performance, enhanced listening, automatic accompaniment and so on. The project has a heavy data-driven component, since we aim to develop deep learning based approaches to music alignment, which require large datasets containing labelled alignments.
Contribution: The overall goal of the project QMUL3 is to develop robust alignment methods which have significant coverage and can adapt to the setting they are being employed in. This involves training of models on different kinds of audio data to improve the coverage, and additional fine-tuning of models based on local context and by leveraging user data. There is a significant requirement of alignment data in these terms to build strong alignment models. The data-driven aspects of this project overlap significantly with the user-driven aspects, since we leverage a combination of labelled data (pre-annotated) and contextual data (from the users' local context). We demonstrated in our paper at EUSIPCO 2020 that our data-driven models which learn frame similarity directly from data perform better than handcrafted features such as chroma-based features or CQT transforms. We employ the MAPS database [3], the Saarland database [4] and the Mazurka dataset [5] for our experiments. Our models outperform traditional methods based on dynamic time warping [6] that rely on handcrafted features, as well as a multi layer perceptron model [7] which learns binary similarity between audio frames. We also demonstrate that salience representations [8] and data augmentation are effective techniques to improve alignment accuracy.

Impact: We foresee a significant impact on both the scientific community and the music market in the long run. This is due to the unprecedented amount of interest generated by automatic music processing tools, coupled with the advancement in artificial intelligence and the massive surge of data available on different platforms; making it possible to build systems which are capable of cutting the costs of human intervention on difficult tasks like transcription and alignment. The research conducted herein could be employed for building

[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.

robust systems to aid online music lessons, 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.

In the future, we will work on modelling structural deviations in music alignment using limited data. Thereafter, we will focus on end-to-end alignment models. A challenge to be faced is the availability of training data comprising manually labelled alignments. We will leverage existing publicly available datasets as well as the data obtained from the Tido platform. Another challenge to be addressed is to optimally deal with the multimodal nature of the datasets and the long-term dependencies present in the inputs. We will attempt to tackle these challenges using novel architectures which are capable of handling long-term dependencies as well as multi-modal input configurations.

3.4 *QMUL4: Drum Sound Query by Vocal Percussion*

This project explores techniques to query drum sounds using vocal imitations as input, usually plosive consonant sounds. The project will output two new datasets for Query by Vocal Percussion (QVP) which could also be useful for other problems in MIR.

The first dataset, recorded and annotated between March and May 2019, is the Amateur Vocal Percussion (AVP) dataset [9]. It comprises a total of 9780 vocal imitations of four types of drums (kick drum, snare drum, closed hi-hat, and opened hi-hat) recorded by 28 participants with fully annotated onsets and labels. In contrast with the previous datasets for QVP, the AVP dataset focuses exclusively on people with little or no experience in beatboxing. It is also directed to query by example applications, allowing algorithms to be trained on several isolated utterances (around twenty-five per label) before they are tested on improvisation recordings. This particular “train-test” organisation of the dataset is also novel and it gets closer to real-life QVP contexts, as the user would first provide example utterances to the algorithm to let it “know” how he/she imitates each drum sound and then test it in realistic beatbox-like improvisations.

The second one, which is expected to be recorded between November 2020 and January 2021, is the Active Learning of Vocal Percussion Style (ALVPS) dataset (tentative name). It would be the consequence of a user study in which around 20 participants would try different active learning algorithms that learn how each of them imitates drum sounds. Therefore, by providing imitations in a sequential way, participants would unknowingly be tuning their system in a dynamic and interactive way. The active learning algorithm would throw strategic drum samples to the participants with the goal of covering its knowledge gaps and, as a result of this approach, fewer input imitations would be necessary to model the user's imitation style. All recordings derived from this process would constitute the ALVPS dataset.

Apart from their use in QVP, these two datasets could be used in other related MIR subfields and problems like rhythm perception and percussive timbre perception among others. They would also be relevant for other audio analysis fields that are closely related to MIR, like speech recognition, and in particular speaker recognition and acoustic phonetic analysis.

3.5 *QMUL5: Adversarial Attacks in Sound Event Classification*

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

[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.

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.

This project uses the following datasets:

- Jamendo dataset: Consists of 93 songs for a total of around 6 hours of music. The dataset represents mainstream music genres. The dataset is split into 61 tracks for training, 16 for validation and 16 for testing.
- FSDKaggle2018: The second dataset used in our work is the FSDKaggle2018 dataset [10] that was released for task 2 of the DCASE 2018 challenge on "General-purpose audio tagging of Freesound content with AudioSet labels". The dataset has 41 labels and it is a single label classification task. The labels range from musical instruments such as snare drum and clarinet to urban sounds such as bus and gunshots. There are roughly 9K audio files in the training data and 1.6K audio files in the test data. The test data is used in our experiments to generate adversarial attacks.

Several data-driven deep learning models are currently used in the project:

- For singing voice detection, the two models we use are the baseline model from [11] and the zero-mean convolution model from [12].
- For FSDKaggle2018, the deep learning models used are from groups that participated in task 2 of the DCASE 2018 challenge [13, 14].

An adversarial attack is a small perturbation added to the input of a machine learning system in order to fool the output. By demonstrating the presence of adversarial attacks in data-driven models we highlight the security risk of deploying data-driven models for public use. More generally, the presence of adversarial attacks highlights properties of machine learning models that are not completely understood.

3.6 *JKU1: Large-Scale Multi-modal Music Search and 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 intuitively explorable in a 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 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, alignment of multiple performances to sheet music for purposes of score-based listening and comparison, and piece identification in unknown recordings. Therefore data-driven aspects in project JKU1 are manifold and play an essential role in the research process.

As for the importance and type of data in this project, we primarily use deep learning methods to learn embeddings directly from the data. Therefore, it is expected that more data will result in better performance and generalisation for our proposed systems. Currently we employ the MSMD dataset [15] to train our

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

network, which consists of a large collection of synthetically generated audio recordings and scores with detailed alignment between note onsets and noteheads thereof, respectively.

The MSMD dataset can be re-rendered for different tempi and instrumentations. Throughout our work we exploited these capabilities by training our system with certain augmentation configurations in order for it to learn tempo-invariant embeddings for the task of audio-to-sheet retrieval, by introducing a soft-attention mechanism to the approach presented in [15]. The motivation is that global and tempo deviations in music recordings might lead to discrepancies between what the system model has seen during training, and the data it will see during test time. This work, the proposed methods and respective experimental results are described in a paper that was published at the 20th International Society for Music Information Retrieval Conference (ISMIR 2019), Delft, The Netherlands [16].

Our current research involves addressing the task of audio-to-score piece identification, that is, retrieving the corresponding document within a collection, given a music piece as input, by exploiting the temporal dependencies between subsequent embeddings. For this use we are exploiting real-word data: commercial audio recordings by various famous pianists of compositions by Mozart, Beethoven and Chopin, as well as the respective scanned images of scores from commercial publishers. This data is also evaluated and described in more detail in [15]. Challenges concerning the data-driven aspects of our project are various, and we see them as opportunities for further research; we then include them in our future work plans. Since one of the targets of this project is to develop retrieval systems which are scalable to large collections, the amount of data available is an essential aspect for us. The aforementioned commercial recordings comprise 193 pieces at this moment, which can be considered far from a large number. Therefore we plan to greatly increase the number of commercial recordings. We also identify the web as providing opportunities for this. Score images can be crawled from the well-established IMSLP Petrucci Music Library, which contains over 400,000 scores; as for audio recordings, we can exploit online video-sharing platforms, such as Youtube, which contains a large number of recordings. Since we do not have the rights associated with these data; we plan also to investigate how to proceed in terms of research dissemination in regard to high-level annotations, which in this case are the mappings of which score document is associated with which recording.

As for the data-driven aspects of our future work, we see great potential in going from synthetic to real-world data during the training stage of the system pipeline. This would allow for better generalisation for our model, which can lead to improved performance in the tasks we have been addressing. This can be done by acquiring (as mentioned in the last paragraph) and annotating multimodal real-world data. In this case, annotations required for preparing the data for training are the correspondences between note events from the audio recordings and score images, which are note onsets in the former, and notehead, measure line, and system coordinates in the latter. We should however say again that we must be careful with data rights, as mentioned before, and investigate, in case we take this path, whether it would be possible to make these annotations publicly available. This would have a significant impact in the MIR community, mainly in the multimodal MIR branch, since there is a lack of openly available annotated multimodal data.

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

The goal of project JKU2 is to develop a live music tracking system that follows an opera performance in real-time along with its respective sheet music. Such a system would be extremely useful for many processes surrounding live opera staging and streaming, including automatic lyrics displays, camera control, or live video cutting. The project requires opera data (audio and/or video recordings, and corresponding PDF scores) but also manual annotations useful to evaluate the tracker. Motivated by the lack of freely available opera datasets, we plan to contribute to the creation of a complete opera dataset including recordings and manual bar-level annotations.

[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.

Data acquisition: As a first step towards opera-to-score alignment, there is a clear need for opera data. So far, no annotated opera dataset is freely accessible, so we had to create our own. As partnership of the project, the Vienna State Opera provided the recordings of two live performances of "Don Giovanni" in the form of two continuous recordings of almost 4 hours each, including the multitrack audio recordings from more than 20 microphones, the video recordings of the main stage, the conductor, and the post-processing result (involving several camera points of view) that has been broadcast by the Opera on their live streaming platform. As a second live performance, we obtained the live recording of "Die Zauberflöte" in a form of an abbreviated version of 2.5 hours played at the Anton Bruckner Universität, including audio and video. We also collected several CD recordings of the two performances (Karajan 1985 and Harnoncourt 1989 for "Don Giovanni"; Karajan 1980 for "Die Zauberflöte"). Full scores of the two mentioned operas, including lyrics, are provided by the Mozarteum Foundation (Stiftung Mozarteum), Salzburg, in PDF format.

Data annotation: To link each recording to its corresponding score, annotations are required. These annotations, also useful to evaluate the accuracy of the tracker, are placed at bar boundaries. For the work "Don Giovanni", a total of 5304 bar-level annotations per performance were created for the four audio recordings. This dataset was partially used in [17] and in [18]. We also plan to annotate the two recordings of "Die Zauberflöte" in the same way.

Another part of this project is to focus on lyric tracking, with the idea of ending with an automatic subtitle system. In this way, we plan to annotate full-length operas at the sentence level in the near future (the lyric cutting at the sentence level has also been provided by the Vienna State Opera for the two Mozart compositions). These annotations could also contribute to the development of an audio-to-lyrics alignment system that is able to detect and track words inside songs. There is also the possibility of extending the topic to opera performances with the constraint of working in real-time.

Variance in manual annotations: The project aims at developing an algorithm that follows the score while the corresponding opera is played online. The accuracy of such a system is on average below 600 ms [18]. This measure corresponds to the mean of the absolute time difference between the alignment provided by the system and the corresponding manual annotations. However, it is useful to remember that an opera involves a hundred instruments and a dozen actors playing on stage at the same time. In that respect, it is difficult to define the ground truth position of the bar annotations precisely. This implies that the manual annotations rely heavily on the choices of the annotator. A previous study of annotations conducted on classical music recordings [19] showed that the median value of standard deviation can reach 70 ms and we believe it would be a much higher figure in opera annotations.

Data publication: For this project, we aim at contributing to the open data situation related to our particular research problem (opera tracking). Because there are no openly available annotated opera corpora for this research, we plan to invest quite some effort into preparing an annotated opera and making it available to the research community, via an open data repository. This concerns both audio and video recordings, and various kinds of manual annotations.

In particular, we have at our disposal the rights to publish the live audio and video recordings of "Die Zauberflöte" with their annotations. Also, we are still in discussions with the Vienna State Opera regarding legal options for publishing the professional recordings from the opera house. These negotiations are currently in progress.

[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.

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

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. Hence, we formulate this problem as an auto-tagging one where the goal is to predict the contextual use of a given track and user. To address this problem, we take an incremental approach where in the first stage we focus on context prediction using only audio content. This data-driven methodology relies on using the audio content of the tracks in contextual playlists to predict their use. Hence, here we rely on data via two aspects: 1) using the playlists and their title to induce their contextual use, 2) using the audio content of these tracks to derive the musical "style" that fits these contextual uses. Since we approach the problem as an auto-tagging task, it is also important to be aware of the progress in this specific domain. Tags are a popular way to categorise music in large catalogues in order to facilitate their exploration and music retrieval on demand [20]. Music tags include different categories such as emotions (sad, happy), genres (rock, jazz), instrumentation-related (guitar, vocals), or listening activities (dance, relax, workout). Traditionally, tags were assigned to music items by humans, either through editors or through crowdsourcing. However, with the expanding availability of online music, there have been also increasing efforts towards developing music auto-tagging models, i.e. systems that do not require to manually annotate the tracks [21]. Music autotaggers are models trained to automatically predict the correct tags for a given music track from the track content. Several models have been proposed that use the audio content, either as raw signal [22, 23, 24] or pre-processed spectrograms [24,25, 23, 26], to predict the appropriate tags.

3.9 *TPT2: Text-Informed Lead Vocal Extraction*

Due to the high complexity of musical audio mixtures, where sound sources are usually highly correlated in time and frequency and may be modified by (non-linear) audio effects, machine learning techniques are required for the separation task. State-of-the art performance on singing voice separation is achieved by deep neural networks [27-29] which are trained in a supervised fashion on datasets of music mixtures with corresponding isolated vocals such as MUSDB18 [30]. In this project this data-driven approach to singing voice separation is applied as well. The focus has been on recurrent neural networks (RNN) [31] which can model the sequential properties of data such as audio and text.

One novelty of the work done within project TPT2 is that a multi-modal approach is applied. This means the task is not only learned on pairs of audio signals (mixture and vocals) but also additional information about the voice signal is exploited. We are particularly interested in exploiting the information inherent in the lyrics. A multi-modal source separation model based on RNNs has been developed and tested with voice activity information. We were able to show that the model indeed exploits the side information and that the separation quality is improved. The work is published in [32]. We also exploited text information with the model in a multi-modal approach to speech-music separation [33]. Work on exploiting lyrics for singing voice separation is currently in progress.

Another novelty of this project is the combination of singing voice separation and automatic lyrics to audio alignment. These tasks have traditionally been performed sequentially: first the singing voice is separated from other instruments and then the lyrics are aligned. Our multi-modal model learns to align the inputs (text and audio) using attention [34] while being trained for the separation task. This means the alignment is learned without supervision. This is relevant because supervised learning would require a dataset of music recordings with aligned lyrics. Such a dataset exists where audio and text are aligned at word level [35] but in order to inform the separation task we require phoneme level alignment. Creating such fine alignments manually is extremely tedious and the unsupervised approach is preferable. The fact that we include the separation objective makes the alignment system more robust to the presence of other instruments in the audio signals. One of the shortcomings of current data-driven methods is that they need task-specific datasets for supervised learning as mentioned above. In this project we show how this constraint can be relaxed by exploiting

[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.

complementary information or learning without supervision. This becomes increasingly important for the development of more robust data-driven approaches: in order to become more robust, they need to be trained on a large amount of data, so that as much complexity and variety as possible are represented. However, labelling vast amounts of data is too expensive.

One possible perspective of future work on data-driven methods is exploring unsupervised approaches to singing voice separation. Semi-supervised approaches might also be explored where the separation task is learned from audio mixture and lyrics pairs. This would allow exploiting much more data than current supervised methods do.

3.10 *TPT3: Muti-modal Music Recording Remastering*

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 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 data-driven approaches to source separation, an approximation of the source parameters can be obtained using non-linear regression techniques where either the clean sources [36, 37] or their corresponding time-frequency masks [38, 39] are used as training target. Thus, the mapping is learned from examples. However, in order to have a good approximation, a considerable amount of training data is needed. These types of data consist not only in the musical mixtures, but also their constitutive isolated sources. Most of the time, the isolated sources are either unavailable or private and involve many copyright problems. In the last years, the appearance of a few publicly available datasets of this kind paved the way to research on data-driven methods for music source separation. However, such datasets account only for a small selection of instruments, usually the most common ones. Moreover, the availability of a large amount of data is not always realized, especially when working on informed source separation where the annotation of the side information is needed. When the side information is the listener's neural response, as in the case of my project, this is particularly true as the acquisition process of such data can be very long and expensive. In such cases unsupervised techniques are still needed.

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).

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 [40, 41], we are proposing a combination of data-driven and knowledge-driven approaches, which consists in generating a synthetic training dataset and subsequently using it to train a neural network. The dataset has the following properties:

- It is generated based on human-defined rules and patterns (exploiting a commercial accompaniment generation software).
- It is categorized into a large number (thousands) of narrowly defined style classes.
- 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 the synthetic dataset.

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

In the 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.

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

Generative models require an immense amount of training data to model the probability distribution of a certain process accurately. Therefore, the project inevitably addresses some aspects related with a data-driven approach, as we necessarily have to build architectures that are scalable to large-scale datasets. This is of particular importance in the case of Generative Adversarial Networks, which perform implicit density estimation and, by nature, they have a mode-seeking behaviour (as opposed to mode-covering models). Therefore, it is important to have large-scale datasets in order to capture variety. As opposed to the field of computer vision, which has seen a huge improvement in the availability and development of large-scale research datasets, the MIP field, still lacks consensus in defining a benchmark dataset for generative modelling of audio. Some of the most remarkable works attempting audio generation with neural networks have made use of solo piano databases [42, 43], music extracted from video-games [44] or synthetically generated single-note audio such as NSynth [45], which is almost the standard for these types of tasks at the moment (although it just covers pitched sounds). Throughout our work, we exploited the NSynth dataset for different experiments, including the comparison of representations for audio synthesis using GANs. For the second paper, we employed a private, non-publicly available dataset, obtained from Sony CSL's private collection that contains a large set of kick drum, snare drum and cymbal samples.

3.13 *UPF1: Facilitating Interactive Music Exploration*

This research is primarily data-driven, as we take the approach of using deep auto-tagging systems to learn the latent space which would be the basis of the music exploration. Data is very important for deep learning systems, as more data usually means better performance and generalization capabilities. Large industrial companies have the upper hand with respect to the data available to them for research, with in-house datasets, while we focus on reproducibility and open science.

Dataset	MTAT	MSD	FMA	MTG-Jamendo
Tracks	5,405	505,216	106,574	55,609
Artists	230	N/A	16,341	3,565
Tags	188	522,366	161	195
Tag groups	No	No	No	Yes
Full tracks	No	No	Yes	Yes
CC-licensed	No	No	Yes	Yes
Bitrate (kbps)	32	104	263	320
Sample rate (kHz)	16	22-44.1	44.1	44.1

Table 1: Popular music auto-tagging datasets, compared to MTG-Jamendo dataset

There are several open auto-tagging datasets: Million Song Dataset (MSD) [46], MagnaTagATune (MTAT) [47], Free Music Archive (FMA) [48], but each of them has its own limitations that are summarized in Table 1 (taken from [49]). For this reason we have introduced the MTG-Jamendo dataset [49] -- the new open music auto-

[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.

tagging dataset. The main differences to the other open datasets is that all audio tracks are available under Creative Commons license in high quality, the tracks are tagged by artists instead of listeners, and the process is curated by Jamendo. Another difference is that the music on Jamendo is mostly created by independent artists, thus it is different from the typical catalogue of popular music.

To learn a semantic latent space, we consider several of the state-of-the-art auto-tagging architectures that have been trained on several datasets. We use the MTG-Jamendo dataset to evaluate and explore the latent spaces. We consider both tag and embedding spaces, which are extracted with the Essentia-TensorFlow library [5] that provides implementations of the following architectures:

- MusiCNN - musically-motivated CNN [51]. It uses vertical and horizontal convolutional filters, pre-trained on MSD and MTAT.
- VGG - architecture from computer vision [52] based on a deep stack of 3x3 convolutional filters adapted for audio [54], also pre-trained on MSD and MTAT.
- VGGish - original implementation of computer vision architecture [52] with the number of output units set to 3087 [53], pre-trained on AudioSet.

Both extracted embeddings and taggrams can be used as latent spaces for exploration. Taggrams are better in the sense that it is easier for users to understand the meaning of the dimensions of the tag space, while embeddings capture lower-level semantics and are not as easy to interpret. However, embeddings might be useful to capture characteristics that are in between raw spectrograms and conventional music labels and tags. We plan to compare different latent spaces and see how they perform in the context of music exploration.

We have also tested the generalization capabilities of the datasets. We trained the MusiCNN model on one dataset top 50 tags and tested on the other one averaging the performance across 7 tags that are common for all 3 datasets. The results in Table 2 show that there are some generalization issues between datasets, thus datasets indeed have internal biases that need to be overcome.

Train \Test	MTAT	MSD	MTG-Jamendo
MTAT	0.95	0.83	0.84
MSD	0.91	0.90	0.84
MTG-Jamendo	0.90	0.82	0.87

Table 2: Cross-dataset evaluation (ROC-AUC)

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

Most of the work conducted until now in project UPF2 is based on data-driven approaches. However, some of the algorithms are designed with an eye on the final user. We will, therefore, present firstly the data-driven work and, in the user-driven section, we will explain the design choices taken with the final user in consideration.

Research focused on loops faces a lack of open reference datasets. Most of the existing work uses private collections which have to be bought from commercial sources, and different works use different collections. There is no unified collection for comparing research on loops, and the freely available datasets either do not

[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.

contain sufficient annotations or lack licenses which allow distribution. To promote open and accessible research on audio loops, we propose a free and distributable database of loops from Freesound 3: FSLD. This dataset contains production-ready loops which are distributed under Creative Commons licenses and can, therefore, be freely shared among the research community and the industry. Part of the dataset has been manually annotated with information about rhythm, tonality, instrumentation and genre; in a similar way as commercially available loop collections are annotated. The annotation service is made public so that the community can work on enlarging the annotations of this collection. We expect this dataset to have an impact on the academic community as it supports further research into several timely research topics which are also of great interest to the industry.

To complement this public dataset in our research, we have also gathered private collections of loops which cannot be shared with other researchers. Loops from Looperman, a community database which does not allow the redistribution of sounds, were collected together with annotations for tempo, instrument, genre and key. Finally, we use a private collection of loops originating from different commercial sources. These large datasets of loops allow us to evaluate our research with several types of collections: a community database which has audio recordings which may or not be loops; another community database focused mainly on loops; and finally, a collection similar to what a music maker would have on their local computer.

Our work into the generation of loops starts with the synthesis of one-shot percussive sounds which we extend for the generation of percussion loops. To this end, we present a deep neural network-based methodology for synthesising percussive sounds with control over high-level timbral characteristics of the sounds. This approach allows for intuitive control of a synthesizer, which was not present in existing work on drum synthesis [55] and which we will further detail in the user-driven section. We use a feedforward convolutional neural network-based architecture, which is able to map input parameters to the corresponding waveform. We use two datasets to evaluate our approach on both a restrictive context (a private collection comprising of commercial kick drum sounds) and on one covering a broader spectrum of sounds (sounds collected and annotated from Freesound), which we use to train our models. We use the input features for evaluation and validation of the model, to ensure that changing the input parameters produces a congruent waveform with the desired characteristics. Finally, we evaluate the quality of the output sound using a subjective listening test.

Following a similar approach to that used in the percussive sound synthesis, we create a system capable of generating drum loops from high-level features. We use the Looperman dataset, and select the loops which only have drum sounds and which have a tempo between 130 -150 BPM. All the loops are time-stretched to 140 BPM and cut to have the length of 1 bar. Besides the timbral features used in the previous work, we extract onset detection functions for each drum sound using the approach of Southall et al. [56]. We evaluate the model synthesis quality through a listening test and the feature coherence between the input and the output by modifying the timbral characteristics of the data in the validation set.

Our work on the automatic classification of instrumental sounds started by analysing the consequences of applying audio effects on instrumental recordings in a state-of-the-art model for this task. We process NSynth [57], a large-scale dataset of one-shot sounds originating from digital synthesisers with several effects such as delay, reverb, distortion and chorus. We then train the model proposed by Pons et al. [58] on either the original or augmented versions of the dataset, where effects are applied. Finally, we evaluate these models in the original and processed versions of the test set. We identify that the accuracy of this algorithm is greatly decreased when tested on sounds where audio effects are applied and see that the augmentation can lead to better classification of unprocessed sounds. As ongoing work, we are using the FSLD along with a convolutional neural network to automatically identify the instrumentation in a loop.

Finally, we proposed two new libraries to enhance the retrieval and the characterisation of sounds. The first one is a JUCE client library which permits easy integration of Freesound in JUCE projects. JUCE is the most used

[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.

library for developing commercial plugins and audio applications. The presented library allows, among other things, to make use of the advanced text and audio-based Freesound search engine, to download and upload sounds, and the retrieval of a variety of sound analysis information (i.e. audio features) from all Freesound content. This enables audio applications to make use of Freesound as their source of audio data, which can lead to an increase in Freesound's popularity.

The second library proposed is TIV.lib, an open-source library for the content-based tonal description of musical audio signals. Its main novelty relies on the perceptually-inspired Tonal Interval Vector [9] space based on the discrete Fourier transform, from which multiple instantaneous and global representations, descriptors and metrics are computed e.g., harmonic changes, dissonance, diatonicity, and musical key. The library is cross-platform, implemented in Python and in the graphical programming language Pure Data, and can be used in both online and offline scenarios for large-scale commercial, scientific, and educational applications.

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

The pioneering works of data-driven version identification (VI) were built upon previous knowledge-driven systems [60, 61], rather than proposing substantial changes in the workflow. The primary focus was to improve scalability by avoiding computationally-expensive local alignment algorithms, and this trend resulted in important advances for building lighter and faster systems [60, 62]. However, such systems had problems with achieving the accuracy standards of their predecessors [63]. Recent data-driven systems that use deep learning techniques pave the way to encapsulate the similarities among versions in ways that are both efficient and accurate [64, 65, 66, 67, 68].

Our first step for a novel data-driven approach was to prepare a dataset. Our motivation for this was rooted in 3 main observations: (1) the publicly-available datasets for VI are small in size, (2) due to copyright laws, only pre-extracted features can be shared publicly, and (3) high-performance deep learning approaches are data-hungry in nature. To address these points, we created the Da-TACOS benchmark [63] and, later, training subsets [65]. The benchmark subset includes 15k songs in total, which made it the largest benchmark set when released. It includes a wide range of pre-extracted features (3 PCP variants, MFCC, and 4 rhythm features) along with metadata. The training subset includes 98k songs for a more limited range of pre-extracted features and metadata. With Da-TACOS, we were able to develop state-of-the-art VI models and evaluate them, and by sharing it publicly, we took a step toward facilitating reproducibility for VI research.

After obtaining a large dataset for model development, our next work was to create a novel deep learning model for VI. The previous deep learning-based VI approaches mostly followed a task-agnostic approach for constructing their network architectures and training strategies [67, 69], using attested design decisions from systems developed in other research communities (e.g., computer vision), which was a common trend until recently. Believing that a task-specific approach would be better suited for our purposes, we started to develop MOVE, musically-motivated version embeddings [64].

MOVE combines the advantages of both knowledge-driven and data-driven design strategies for VI. The network architecture includes techniques for achieving transposition and structure invariances, which are common challenges in VI literature. The training strategies include data augmentations specifically addressing the commonly observed changes in musical characteristics among versions (i.e., pitch transpositions and tempo and micro-timing variations). The network is trained with a metric learning approach to address the scalability issue that prevents successful VI systems from being used in large-scale use cases.

The best performing configuration of MOVE has achieved a relative 12% increase in the mean average precision metric compared to its closest competitor on the Da-TACOS benchmark set. Moreover, while MOVE takes only

[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.

a few minutes to estimate all pairwise similarities in Da-TACOS, its closest competitor takes a few weeks. These results have proven the potential of data-driven systems for bridging the gap between accuracy and scalability, as well as the importance of incorporating task-specific knowledge into network design.

Although a substantial step toward improving scalability has been made with MOVE, the best performing configuration was still cumbersome to be used in industrial applications. To improve scalability even further, we have studied model compression techniques for obtaining smaller embedding vectors that encode songs. By reviewing existing techniques and proposing new ones, we investigated the effectiveness of embedding distillation methods for VI [65].

Specifically, embedding distillation techniques aim to find a model that outputs smaller embedding vectors by taking advantage of a pre-trained model. In other words, if a pre-trained model produces large embedding vectors but is strong in performance, we aim to find a new model that mimics the performance of the pre-trained model while producing smaller embedding vectors. We have investigated neural network pruning and knowledge distillation techniques and proposed a technique called latent space reconfiguration, which uses the strong priors of a pre-trained feature extractor to reconfigure the latent space where the embeddings lie.

Using embedding distillation techniques, our proposed model Re-MOVE has improved the accuracy of MOVE by 3% while producing 98% smaller embedding vectors. This is an important advancement for scalability since smaller embeddings require less storage space and allow for faster retrieval.

Both MOVE and Re-MOVE propose a number of reusable insights that are aimed to facilitate building better systems for VI. However, due to the input representation both models use, the similarities among versions are investigated only from a tonal perspective, which cannot provide a complete picture considering the multi-dimensional nature of music. To address this issue, we have investigated methodologies that combine models trained for capturing harmonic and melodic similarities, in a collaboration with researchers from IRCAM [66].

For understanding the pros and cons of systems that use melodic and harmonic input representations, we have experimented with different models and features. The initial results showed that the models that use the harmonic representations, as MOVE and Re-MOVE, yield better performances for VI compared to the ones that use melodic features. However, with further experiments, we have shown that by combining the obtained embedding vectors of two models that use different types of inputs, we were able to outperform both models substantially. Although the resulting system is considerably larger than MOVE or Re-MOVE alone, the increase in performance indicates that incorporating various musical dimensions for investigating similarities among versions is a research direction that is promising and undervalued.

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

[2] 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, October, 2019.

[3] Valentin Emiya, Nancy Bertin, Bertrand David, and Roland Badeau. MAPS: A piano database for multipitch estimation and automatic transcription of music. 2010.

[4] Meinard Müller, Verena Konz, Wolfgang Bogler, and Vlora Arifi-Müller. Saarland music data (smd). In Proceedings of the international society for music information retrieval conference (ISMIR): late breaking session, 2011.

[5] Craig Stuart Sapp. Comparative analysis of multiple musical performances. In ISMIR, pages 497-500, 2007.

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

[7] Ozgür Izmirli and Roger B Dannenberg. Understanding features and distance functions for music sequence alignment. In ISMIR, pages 411-416. Citeseer, 2010.

[8] Rachel M Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan Pablo Bello. Deep salience representations for f0 estimation in polyphonic music. In International Society for Music Information Retrieval (ISMIR), pages 63-70, 2017.

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

[10] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, and X. Serra, "Generalpurpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), (UK), pp. 69-73, 2018.

[11] J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks.," in ISMIR, pp. 121-126, 2015.

[12] J. Schlüter and B. Lehner, "Zero-mean convolutions for level-invariant singing voice detection.," in ISMIR, pp. 321-326, 2018.

[13] T. Iqbal, Q. Kong, M. D. Plumbley, and W. Wang, "General-purpose audio tagging from noisy labels using convolutional neural networks," in Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), (UK), pp. 212-216, 2018.

[14] I.-Y. Jeong and H. Lim, "Audio tagging system for dcase 2018: focusing on label noise, data augmentation and its efficient learning," tech. rep., Tech. Rep., DCASE2018 Challenge, 2018.

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

[16] Stefan Balke, Matthias Dorfer, Luis Carvalho, Andreas Arzt, and Gerhard Widmer. 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, 2019.

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

[17] Charles Brazier and Gerhard Widmer. 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), pages 371–377, Turin, Italy, 2020.

[18] Charles Brazier and Gerhard Widmer. Addressing the Recitative Problem in Real-Time Opera Tracking. In Under review.

[19] Thassilo Gadermaier and Gerhard Widmer. 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), pages 769–775, Delft, The Netherlands, 2019.

[20] Paul Lamere. Social tagging and music information retrieval. Journal of New Music Research, 37:101- 114, 2008.

[21] Thierry Bertin-Mahieux, Douglas Eck, Francois Maillet, and Paul Lamere. Autotagger: A model for predicting social tags from acoustic features on large music databases. Journal of New Music Research, 37(2):115-135, 2008.

[22] Taejun Kim, Jongpil Lee, and Juhan Nam. Sample-level cnn architectures for music auto-tagging using raw waveforms. In Proceedings of the 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2018.

[23] Jordi Pons Puig, Oriol Nieto, Matthew Prockup, Erik M Schmidt, Andreas F Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale. In Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR, 2018.

[24] Minz Won, Sanghyuk Chun, and Xavier Serra. Toward interpretable music tagging with self-attention. arXiv preprint arXiv:1906.04972, 2019. [17] Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks. arXiv preprint arXiv:1606.00298, 2016.

[25] Keunwoo Choi, Gyorgy Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[26] Minz Won, Sanghyuk Chun, Oriol Nieto, and Xavier Serra. Data-driven harmonic filters for audio representation learning. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 536-540. IEEE, 2020.

[27] F.-R. Stoter, S. Uhlich, A. Liutkus, and Y. Mitsufuji. Open-unmix - a reference implementation for music source separation. Journal of Open Source Software, 2019.

[28] Naoya Takahashi, Nabarun Goswami, and Yuki Mitsufuji. MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation. In International Workshop on Acoustic Signal Enhancement, pages 106-110. IEEE, 2018.

[29] Alexandre Defossez, Nicolas Usunier, Leon Bottou, and Francis Bach. Demucs: Deep extractor for music sources with extra unlabeled data remixed. arXiv preprint arXiv:1909.01174, 2019.

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

[31] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[32] Kilian Schulze-Forster, Clement Doire, Gaël Richard, and Roland Badeau. Weakly informed audio source separation. In 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pages 273{277. IEEE, 2019.

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

[33] Kilian Schulze-Forster, Clement Doire, Gaël Richard, and Roland Badeau. Joint phoneme alignment and text-informed speech separation on highly corrupted speech. In 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020), 2020.

[34] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[35] Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters. Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. arXiv preprint arXiv:1906.10606, 2019.

[36] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain. arXiv preprint arXiv:1911.13254, 2019.

[37] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. arXiv preprint arXiv:1806.03185, 2018.

[38] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. Open-unmix - a reference implementation for music source separation. 2019.

[39] Yi Luo, Zhuo Chen, John R Hershey, Jonathan Le Roux, and Nima Mesgarani. Deep clustering and conventional networks for music separation: Stronger together. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 61–65. IEEE, 2017.

[40] Ondřej Cífka, Umut Şimşekli, Gaël Richard. “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. doi:10.5281/zenodo.3527878.

[41] 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).

[42] Aaron van den Oord et al. “WaveNet: A Generative Model for Raw Audio". In: The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016.

[43] Curtis Hawthorne et al. “Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset". In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.

[44] Edith Law and Luis von Ahn. “Input-agreement: a new mechanism for collecting data using human computation games". In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009. 2009, pp. 1197-1206.

[45] Jesse Engel et al. “Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders". In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. 2017, pp. 1068-1077.

[46] Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B. & Lamere, P. The million song dataset. In Klapuri, A. & Leider, C. (eds.) Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011, 591–596 (University of Miami, 2011).

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

[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.

[48] Defferrard, M., Benzi, K., Vandergheynst, P. & Bresson, X. FMA: A dataset for music analysis. In Cunningham, S. J., Duan, Z., Hu, X. & Turnbull, D. (eds.) Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017, 316–323 (2017).

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

[50] Alonso-Jiménez, P., Bogdanov, D., Pons, J. & Serra, X. Tensorflow audio models in essentia. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 266–270 (2020).

[51] Pons, J. & Serra, X. musicnn: Pre-trained convolutional neural networks for music audio tagging (2019). 1909.06654.

[52] Simonyan, K. & Zisserman, A. Very deep convolutional networks for largescale image recognition. In Bengio, Y. & LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015).

[53] Choi, K., Fazekas, G. & Sandler, M. B. Automatic tagging using deep convolutional neural networks. In Mandel, M. I., Devaney, J., Turnbull, D. & Tzanetakis, G. (eds.) Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016, 805–811 (2016).

[54] Hershey, S. et al. CNN architectures for large-scale audio classification. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017, 131–135 (IEEE, 2017).

[55] Pritish Chandna, Merlijn Blaauw, Jordi Bonada, and Emilia Gómez.Wgansing: A multi-voice singing voice synthesizer based on the wasserstein-gan. In 2019 27th European Signal Processing Conference (EUSIPCO), pages 1–5. IEEE, 2019.

[56] Soonbeom Choi, Wonil Kim, Saebyul Park, Sangeon Yong, and Juhan Nam. Korean singing voice synthesis based on auto-regressive boundary equilibrium gan. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7234–7238. IEEE, 2020.

[57] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. arXiv preprint arXiv:1802.04208, 2018.

[58] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. arXiv preprint arXiv:1902.08710, 2019.

[59] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1068–1077. JMLR. org, 2017.

[60] T. Bertin-Mahieux and D. P. W. Ellis, “Large-scale cover song recognition using the 2D Fourier Transform magnitude,” in Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR), 2012, pp. 241–246.

[61] S. Ravuri and D. P. W. Ellis. Cover song detection: from high scores to general classification. In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2010, pp. 65–68.

[62] E. J. Humphrey, O. Nieto, and J. P. Bello. Data driven and discriminative projections for largescale cover song identification. In Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR), 2013, pp. 4–9.

[63] F. Yesiler, C. Tralie, A. Correya, D. F. Silva, P. Tovstogan, E. Gómez, and X. Serra. 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.

[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.

[64] 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.

[65] 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 (in print).

[66] 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 (in print).

[67] Z. Yu, X. Xu, X. Chen, and D. Yang. Temporal pyramid pooling convolutional neural network for cover song identification. Proceedings of the International Joint Conference on Artificial Intelligence, 2019, pp. 4846–4852.

[68] C. Jiang, D. Yang, and X. Chen. Learn a robust representation for cover song identification via aggregating local and global music temporal context. In Proc. of the International Conference on Multimedia and Expo (ICME), 2020, pp. 1–6.

[69] G. Doras and G. Peeters. Cover detection using dominant melody embeddings. In Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR), 2019, pp. 107–114.

[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.