TROMPA

TROMPA: Towards Richer Online Music Public-domain Archives

Deliverable 3.2

Music Description

Grant Agreement nr	770376
Project runtime	May 2018 - April 2021
Document Reference	TR-D3.2-Music Description v2
Work Package	WP3 - Automated Music Data Processing and Linking
Deliverable Type	Report
Dissemination Level	PU- Public
Document due date	30 April 2020
Date of submission	30 April 2020
Leader	UPF
Contact Person	Aggelos Gkiokas (aggelos.gkiokas@upf.edu)
Authors	Aggelos Gkiokas (UPF), Emilia Gomez (UPF), Helena Cuesta (UPF), Olga Slizovskaia (UPF), Juan Gomez-Cañón (UPF), Lorenzo Porcaro (UPF)
Reviewers	Cynthia Liem (TUD)

Executive Summary

This aim of **Task 3.2 Music Description** is to develop and integrate technologies for the automatic description of the majority of the musical data described and collected in **T3.1 Music Description**. This deliverable is the second and final version of the Deliverable 3.2 - Music Description, part of Work Package 3. The goals of this task is to compute music descriptors from the repertoire that are defined in Deliverable 3.1 Music Resource Preparation. These descriptors were chosen based on the needs of the use cases and state-of-the-art methods will be used. The goals of this task can be summarized to:

- Apply and evaluate existing state-of-the-art music description methods in the domain of classical music and focusing on the target repertoires.
- Develop new and expand existing methods for music description tailored to the target repertoires.
- Facilitate the use cases with robust descriptors.
- Contribute data descriptors to TROMPA repositories for enrichment of public domain repositories and facilitate future musicological or other research.
- Contribute data (descriptors) and algorithms to open repositories such as AcousticBrainz and Zenodo in order to facilitate future research on music processing.

We used and further developed various existing state-of-the-art methods/libraries for the extraction of descriptors. This task is more focused on the extraction of **audio** descriptors, however we also consider **symbolic scores,** and **videos**. We start from the **low level audio** descriptors, such a spectral and cepstral frame based descriptors. At next we will present **mid-level audio** descriptors, such as information related to **rhythm** (beat locations, tempo), the **tonality** (key, chords) and **music performance** (for choir singing: intonation, tuning). Next we present higher level music descriptors, such as **music similarity** and **emotion classification**. These descriptors are more related to human notion about music and must be considered more subjective from the mid-level features. At next we present the descriptors derived from the **symbolic** representations of music (MEI, MusicXML, MIDI) , and finally we will describe descriptors derived from **music videos** (concerts, rehearsals etc). These descriptors fuse information from both audio and video, and thus are considered as **multi-modal** descriptors.

Essentia [1] will be used as the main framework for extracting state-of-the-art audio features. Apart from Essentia, we will deploy other state-of-the-art methods of music description. These methods will be appropriately selected in order to meet TROMPA use case requirements, and will be adapted and further developed. Whenever possible, these methods will be contributed to Essentia, otherwise will be provided and integrated as individual open libraries. The methods that will be developed and the descriptors that we will be extracted are the following:

- **Low-level audio descriptors**: We will use Essentia as the tool to extract these descriptors which among others are **band energies** (bark, mel, ERP), **cepstral descriptors** (mfcc, bfcc), **spectral moments** and **time domain** features (envelope, ZCR)
- **Harmonic-Tonal descriptors**: Include various representations such as the chromagram, Harmonic Pitch Class Profile (HPC), chord descriptors, key, tuning. These descriptors can be potentially use in all cases that involve audio files. To this end, we have identified the following potential use of the rhythm descriptors:
  - **Music Enthusiasts**: Harmonic/Tonal descriptors can be used to facilitate a music recommendation/music similarity engine..

- ➢ **Instrument Players**: Harmonic/Tonal descriptors can be used to analyze existing or new performances of instrument players.
- ➢ **Choir Singers:** As in the instrument players use case, we can use tonal descriptors (e.g. tuning) to characteristics of choir singing performances.
- ➢ **Music Scholars:** The chord descriptors can facilitate musicological research, e.g. retrieving works with similar chord progressions.

❖ **Rhythm descriptors**: Rhythm descriptors contain explicit information about the rhythmic content of a music piece, such as tempo, beats, tempo curves (tempo fluctuations), time signature, rhythm tags and meter and can be potentially use in all cases that involve audio files:

- ➢ **Music Enthusiasts**: Rhythm descriptors can be used to facilitate a music recommendation/music similarity engine.
- ➢ **Instrument Players**: Rhythm descriptors such tempo fluctuations can be used to analyze existing or new performances of instrument players.
- ➢ **Choir Singers:** As in the instrument players use case, we can use rhythm descriptors to analyze rhythm characteristics of choir singing performances.

Apart the use cases the main contribution of this subtask is that we will run the rhythm analysis algorithms on all TROMPA data referred to the CE in the context of enrichment and create new reusable data for future scientific/musicological research.

❖ **Singing voice analysis**: Singing voice analysis involves extracting information about the vocals of a music piece. Singing voice may appear in music pieces in several ways: accompanied or a cappella, solo or ensemble, e.g. choirs. Many expressive properties of singing voice can be extracted, including (but not limited to) **pitch curves** (curves showing the evolution of the fundamental frequency in time), **intonation** (accuracy of pitch in singing), **degree of unison** (the agreement between all the voice sources) and **synchronization**. This particular task is primarily related to the Choir singers use case. Moreover, we contribute to the state-of-the-art in multi-pitch estimation, building new models trained using choral singing data and we will develop a framework capable of extracting intonation descriptors given an audio input and an associated score. In addition, and taking advantage of the human-annotated data that will be gathered in the scope of TROMPA, we contribute to the state-of-the-art of singing performance rating, combining low-level and pitch descriptors with annotations.

❖ **Music Recommendation:** We develop music recommendation methods based on an hybrid system which makes use of both content- (e.g. audio), context- (e.g. tags) and collaborative filtering methods to provide recommendations in the context of the Music Enthusiasts use-case. In particular, recommendations are tailored using emotion information, framing our method within the category of *Psychologically-inspired music recommendation*.

❖ **Emotion Tag Annotation**: In the context of the TROMPA project, we will make use of the categorical approach to annotate musical excerpts with emotion ratings. We will study the agreement in emotion annotation and considering personalized models. Since we aim at using public domain music, we focus on the annotation of choral music. Through the use of the Contributor Environment, we will be able to connect to the Muziekweb website and retrieve excerpts of this type of music for annotation. In this context, our approach is twofold:

- ➢ Analyze the agreement of annotations by the music enthusiasts that participate in our application.

- ➢ Use these annotations to improve the performance of emotion recognition algorithms through personalization methods, i.e. with active learning.

❖ **Symbolic descriptors**: Symbolic descriptors will be used in facilitating tasks related to the **automatic assessment of music pieces** for the choir singers and the instrument players use cases. Moreover we will extract baseline symbolic descriptors that can be potentially exploited in other use cases, such as the music scholars use case. They are based on conventional MIDI processing libraries such as MIDI Toolbox [19] and Pretty-Midi[1]. Apart the explicit use of these features for the use cases or research includes

- ➢ **Rhythm analysis from symbolic streams**: We develop a method for automatic extracting rhythm information (beats, tempo) from symbolic MIDI streams [Ref Dafx].
- ➢ **Difficulty assessment of choir and pieces**: These symbolic descriptors will be used to conduct related to estimating the difficulty of a choir singing a music piece based on its score (Part of **Task 5.4 - Music Performance Assessment**).

❖ **Video descriptors**: Video descriptors are aimed to provide additional semantic information related to video recordings of musical performances. In the scope of the project, we will focus on the following tasks revealing the potential of using video data:

- ➢ **Video tagging**: general-purpose video tagging in the domain of musical performances, providing frame-level and video-level labels from a pre-defined ontology.
- ➢ **Musical instrument detection:** providing a position of an object in a form of a rectangular bounding box.
- ➢ **Automatic object segmentation:** localizing objects from a pre-defined ontology as an enclosed free-form area at a frame level.

Video descriptors will be mainly used for enrichment of TROMPA data irrespectively of any of the use cases. However we foresee that such descriptors can be potentially used in many use cases which involve video such as:

- ➢ **Instrument Players**: showing the fingering charts upon a video recording.
- ➢ **Choir Singers**: counting the number of singers.
- ➢ **Music Enthusiasts**: providing a more detailed transcription for music performances, instrument-to-instrument navigation in music videos, highlighting playing/non-playing instruments in videos.

In order to automate the extraction of the descriptors, all of them will be integrated on to the **TROMPA Processing Library.** Each descriptor extraction program is represented in the CE database as a set of nodes. The definitionThese nodes are related to items such as the software that implements the algorithm, the algorithm (a software can contain more than one algorithm) or the parameters of it. In particular, for integrating to the CE, each algorithm/software has to:

- ❖ Subscribe to the CE through a websocket. This functionality offers the possibility to handle user requests for algorithms in real-time.
- ❖ Frequently check for tasks by querying the CE.
- ❖ Run a task and update the status to the CE of the algorithms running (e.g. ‘running’, ‘completed’, ‘failed’).

[1] https://github.com/craffel/pretty-midi

Apart from the representation of the algorithms on the CE, for each algorithm the following has to be decided:

- **Software Specifications**: Specify the software, OS and other software dependencies.
- **Hardware Specifications:** Specify hardware requirements
- **Physical Location:** Decide on which computer(s) the algorithm should be executed.
- **Arguments:** Specify the parameters of the algorithm and the arguments used
- **Descriptor Dependencies:** Specify if it requires other descriptors: the extraction of one descriptor can be based on other descriptors (sequential processing).
- **Storage:** Where the descriptions are stored (CE, external repositories)

It has to be mentioned, that apart from the use cases, these descriptors will be used to enrich the TROMPA content for **future research** and **future potential** use cases beyond the end of the project. By the end of TROMPA these descriptors will be deposited in various repositories linked with the Contributor Environment as contributions of TROMPA to public domain musical archives (e.g. automatic annotations, low-level descriptors). Moreover, since all algorithms will be "packaged" as components of the **TROMPA Processing Library:**

- Use cases will be "agnostic" of where and how the algorithms are run and can only make requests to the CE to trigger the appropriate algorithms and receive the descriptors needed.
- Algorithms will be potentially available to be re-used by future pilots/applications after the end of the project.

Although this is the final version of this deliverable, scientific research on the topics described in this deliverable will be continued until the end of the project. Any further scientific contributions such as scientific publications or datasets will be published on the corresponding webpage[2] of our website and reported on the periodic reports.

[2] https://trompamusic.eu/research/results/publications

Version Log		
#	Date	Description
1.1	6 April 2020	1st draft version released for internal review
1.2	20 April 2020	Revised version after internal review
2.0	30 April 2020	Final version

Table of Contents

1. Introduction

This aim of **Task 3.2 Music Description** is to develop and integrate technologies for the automatic description of the majority of the musical data described and collected in **T3.1 Music Description**.

1.1 Outline

In the rest of the Introduction section we will briefly describe the score of this deliverable and the goals of Task 3.2. Section 2 is the main part of this deliverable, describing the technologies that have been developed and the relation with the TROMPA use cases and repertoire. In Section 3 we provide some information related to the integration with the TROMPA Infrastructure and the Contributor Environment. Section 4 concludes this deliverable by summarizing the main contributions of this deliverable/task.

1.2 Scope

This deliverable is the second and final version of the Deliverable 3.2 - Music Description, part of Work Package 3. The current version of the deliverable is an extension of the first version and we follow the same structure with it. For reasons of **clarity** and **readability**, whenever needed we keep some of the text of the first version **the same** instead of omitting it, while in other cases **we only report the new advancements** to reflect the progress of the Task from M10 (1st version) to this version (M24).

The aim of this task is to provide music descriptors at various modalities and levels of analysis for the target repertoires in order to facilitate the use cases and enrich TROMPA content. Although this task primarily deals with audio, we also consider symbolic and video music sources. These target repertoires are described in **Deliverable D3.1 Music Resource Preparation**[3], and they are jointly submitted on M18 of the project.

1.3 Task goals

The goals of this task is to compute music descriptors from the repertoire that are defined in Deliverable 3.1 Music Resource Preparation. These descriptors were chosen based on the needs of the use cases and state-of-the-art methods will be used. The goals of this task can be summarized to:

- Apply and evaluate existing state-of-the-art music description methods in the domain of classical music and focusing on the target repertoires.
- Develop new and expand existing methods for music description tailored to the target repertoires.
- Facilitate the use cases with robust descriptors.
- Contribute data descriptors to TROMPA repositories for enrichment of public domain repositories and facilitate future musicological or other research.
- Contribute data (descriptors) and algorithms to open repositories such as AcousticBrainz and Zenodo in order to facilitate future research on music processing.

[3] https://trompamusic.eu/deliverables/TR-D3.1-Data_Resource_Preparation_v1.pdf

2. Music Descriptors

2.1 Overview and Relation to the Use Cases

In this section we describe which music descriptors we consider, a corresponding technical/scientific description, their relation to the use cases and the TROMPA data enrichment as well as with a reference to existing open source repositories and publications. We used and further developed various existing state-of-the-art methods/libraries for the extraction of descriptors. This task is more focused on the extraction of **audio** descriptors, however we also consider **symbolic scores,** and **videos**. A summary of the descriptors for each of the data types and the corresponding use cases are summarized in Table 2.1.

In the next sections we provide details of the music description methods as done in the previous version of this deliverable, and we will focus on the advancements since then. In Section 2.2 we will describe the audio descriptors and we will start from the **low level audio descriptors**, such a spectral and cepstral frame based descriptors. Next we present mid-level audio descriptors. The term mid-level is derived from the fact that these descriptors contain objective audio and music properties of the signal, such as information related to **rhythm** (beat locations, tempo), the **tonality** (key, chords) and **music performance** (for choir singing: intonation, tuning). Compared to the low-level descriptors, the process of extracting these descriptors is not always a pipeline of calculations, but they are rather more complex models based on Machine Learning, such statistical models and neural networks. Next we will present higher level music descriptors, such as **music recommendation** and **similarity**, and **emotion classification**. These descriptors are more related to human notion about music and must be considered more subjective from the mid-level features. In Section 2.3 we will present the descriptors derived from the symbolic representations of music (MEI, MusicXML, MIDI), and in Section 2.4 we will describe descriptors derived from music videos (concerts, rehearsals etc). These descriptors fuse information from both audio and video, and thus are considered as multi-modal descriptors.

Data type	Descriptors	Related Use Cases
Audio files		
	Low level audio descriptors (baseline features such as spectral features, mfccs etc)	Music Enthusiasts, Choir Singers, Instrument Players
	Harmonic Descriptors (Chroma vectors, Pitch Class Profile)	Music Enthusiasts, Choir Singers, Instrument Players
	Rhythm descriptors (tempo, beats, time signature, meter)	Music Enthusiast, Instrument Players, Choir Singers
	Singing voice analysis descriptors	Choir singers
	Audio Similarity (content based audio similarity, timbre similarity, rhythm similarity, melodic similarity)	Music Enthusiasts, Music Scholars, Instrument players

	Mood/Emotion Classification	Music Enthusiasts
Symbolic Scores		
	Low level descriptors (Note densities, inter-onset-intervals, time signature)	Choir Singers, Instrument Players, Music Scholars
	Difficulty Assessment of a Music Piece	Choir Singers, Instrument Players
Videos		
	Counting the number of singers in a choir	Choir Singers
	Musical instrument recognition	Choir Singers, Orchestras
	Cross-modal singing conversion	Choir Singers
	Audio-visual source separation	To be determined

Table 2.1 Summary of the descriptors and the corresponding use cases.

2.2 Audio Descriptors

Essentia [1] is used as the main framework for extracting state-of-the-art baseline audio features. Essentia is a C++ library with Python bindings for audio analysis, description, and synthesis. The library contains a collection of various algorithms which implement standard digital signal processing blocks, statistical characterization of data, and a large set of spectral, temporal, tonal and high-level music descriptors. Essentia is cross-platform and focuses on optimization in terms of robustness, computational speed and low memory usage, which makes it suitable in the context of TROMPA.

Apart from Essentia, we deploy other state-of-the-art methods of music description. These methods will be appropriately selected in order to meet TROMPA use case requirements, and will be adapted and further developed. Whenever possible, these methods will be contributed to Essentia, otherwise will be provided and integrated as individual open libraries.

2.2.1 Low Level and Harmonic/Tonal Audio Descriptors

In this section we will describe methods for common low level audio descriptors. Most of these descriptors are calculated on overlapping time frames, usually from the spectrogram of these frames. The size of the frames are usually between 20 - 100 ms, with an overlap of one quarter to half of the frame size. The time domain descriptors can be calculated either over the whole audio (i.e. envelope) or in a frame basis, as the spectral descriptors. We will use Essentia [1] as the tool to extract these descriptors. A summary of these descriptors is presented in the following list:

- **Band energies**: Computes the energy of various types of spectral band scales from the audio signal. The band scales implemented are:
  - **Bark Scale**: Computes the energy and magnitude on the Bark scale
  - **Mel-frequency scale**: computes the energy and magnitude on the Mel scale

- ➢ **ERP scale**: computes the energy and magnitude on the ERP (or Gammatone) scale.

❖ **Cepstral Descriptors**: Cepstral descriptors are derived when computing the Discrete Fourier Transform on the energies of a band scale.:

- ➢ **BFCC:** Cepstral coefficients on the Bark scale spectral energies
- ➢ **MFCC**: Cepstral coefficients on the Mel-scale spectral energies
- ➢ **GFCC**: Cepstral coefficients on the Gammatone-scale spectral energies

❖ **Other spectral descriptors**:

- ➢ **LPC**: The standard Linear Predictive Coefficients.
- ➢ **Spectral Roll-off**: The roll-off frequency is the frequency under which a certain percentage of the total energy of the spectrum is contained.
- ➢ **Spectral Contrast**: Estimates the strength of the spectral peaks, valleys and their differences.
- ➢ **Spectral Flatness**: Is the ratio between the geometric mean and the arithmetic mean.
- ➢ **Spectral Moments**: Standard statistical moments of the spectrum (centroid, skewness etc)

❖ **Time Domain Descriptors:**

- ➢ **Envelope:** It applies a non-non-symmetric lowpass filter on a signal to extract its envelope.
- ➢ **Envelope Flatness**: The envelope flatness is the ratio between the threshold (low threshold) in the envelope that has the 20% of the values underneath and the threshold (high threshold) in the envelope that has the 95% of the values underneath.
- ➢ **Attack TIme**: The attack time is defined as the duration from when the sound becomes audible to when it reaches its maximum intensity. This descriptor is computed for onsets.
- ➢ **Zero Crossing Rate (ZCR)**: The number of sign changes between consecutive values of signal values divided by the length of the signal. Signals with higher ZCR are more likely to be more noisy.

2.2.2 Harmonic - Tonal Descriptors

Summary

Harmonic/Tonal descriptors contain information about the harmonic/tonal content of a music signal. We consider Essentia for computing various descriptors including:

❖ **Chromagram**: The chromagram representation of the music signal. It is a spectrogram-like representation that is focused on the frequencies of the 12 western chromatic scale.

❖ **Harmonic Pitch Class Profile (HPCP)**: A 12-dimension vector representing the salience of the tones of the 12 tones of the western chromatic scale.

❖ **Chords**: Given the HPCP the chord progression of an audio file will be extracted.

❖ **Chord Descriptors**: Given the chord progression of a music piece, the following descriptors are calculated:

- ➢ **Chord Histogram**: The normalized histogram of chords.
- ➢ **Chord Change Rate**: The rate at which the chords change.
- ➢ **Dominant Chord**: The most frequent chord in the progression.
- ➢ **Dominant Scale**: The scale of the most dominant chord.

- **Key**: Given the HCPC, the dominant key (major/minor keys) is calculated.
- **Dissonance:** The sensory dissonance of an audio signal given its spectral peaks.
- **Tuning**: The tuning frequency of an audio excerpt.

Adaptation for TROMPA

These descriptors will be extracted and evaluated in the context of TROMPA use cases. However we do not intend to further investigate and research the development of these descriptors.

Contribution to the State-of-the-Art

We contribute to the SoA in terms of computing the harmonic/tonal descriptors described in large amounts of western classical music data. This data can be used for further research after the end of the project.

Data requirements

Since we use the built-in engines of Essentia, there are no data requirements for these descriptors.

Human Annotations / Human Data

We didn't collect any human annotation or other related data regarding these descriptors. However these descriptors can be used to release new datasets to facilitate research in automatic music description of the tonal/harmonic content.

Data output format

Descriptor	Representation
Chromagram, Harmonic Pitch Class Profile and other chord descriptors	One binary file per audio file/feature, containing an NxK matrix (Time x Frequency float values)
Chord Progression	One text file per audio file, containing pairs of (time, chord) for storing the chord progressions.
Key signature	One text file per audio file, containing pairs of (time, key signature) for storing key signatures that may change in a single file.
Tuning	One text file per audio file, containing the tuning frequency and the distance to 440 Hz in cents.

Table 2.2. Harmonic descriptors representation.

Relation to the use cases

These descriptors have not been explicitly used yet in any of the use cases. However, according to the development of the use cases, we have identified the following potential use:

- **Music Enthusiasts**: Harmonic/Tonal descriptors can be used to facilitate a music recommendation/music similarity engine and to search for music based on harmonic properties (e.g. chord progressions, key signature, major/minor).
- **Instrument Players**: Harmonic/Tonal descriptors can be used to analyze existing or new performances of instrument players.
- **Choir Singers:** As in the instrument players use case, we can use tonal descriptors (e.g. tuning) to characteristics of choir singing performances.
- **Music Scholars:** The chord descriptors can facilitate musicological research, e.g. retrieving works with similar chord progressions.

2.2.3 Rhythm Descriptors

Summary

In this section we present methods used to automatically extract rhythm descriptors from audio files. Rhythm descriptors contain explicit information about the rhythmic content of a music piece, which can be summarized to the following properties:

- **Beat locations**: The positions of the beats in a music piece.
- **Tempo**: The tempo value in Beats per Minute (BPM). This value can be either an overall estimation on a music piece (assuming that there are no tempo changes in the piece), or computed as a changing value over time.
- **Tempo curves**: Curves that show the evolution of a music piece in time. This is a very important descriptor, since one the main characteristics of classical music is the variations of tempo within a music piece.
- **Time signature**: The time signature of a music piece (e.g. ¾, ⅞)
- **Meter tracking**: In addition to the extraction of beat locations, meter tracking includes to the estimation of the time signature, as well as the position of each meter (downbeat positions)
- **Rhythm tags**: We can extract rhythm tags based on classification strategies (e.g. tags related to rhythm style, speed)

Method Descriptions

We further develop existing state-of-the-art methods of rhythm processing that are based on Long Short Term Memory (LSTM) networks [3] and Convolutional Neural Networks (CNN) [4]. We exploit the use of unlabelled data in an unsupervised schema using Autoencoders (AE) and multi-task learning schemas for related tasks such as tempo estimation, onset detection and beat tracking.

Adaptation for TROMPA

Since we use Machine Learning and Deep Neural Networks for rhythm analysis tasks, the main adaptation to TROMPA is to use classical music derived from the repositories that we consider for training and evaluating our models.

Contribution to the State-of-the-Art

Automatically tracking the rhythm of western classical music is a very challenging task. There are many properties of classical music such as tempo fluctuations, soft onsets, absence of percussive

instruments and changes in the time signature that make rhythm analysis difficult. By taking advantage of the large amount of data available in TROMPA, the opportunity to gather human annotations, and by using modern Machine Learning techniques focusing on LSTMs and CNNs we will contribute to the State-of-the-Art with novel methods based on rhythm analysis focused on classical music. Moreover we will contribute to new datasets as well as new annotation methods. To sum up, our main contributions to the state-of-the-art on rhythm analysis are:

- **Developing** and **evaluating** new methods for rhythm analysis focused on classical music.
- Contributing **new datasets** for scientific research on rhythm analysis
- Contribute to an efficient method of semi-automated annotation of beats using the **Annotator Tool (Task 5.5)** combined with the output of several algorithms.

Data requirements

The baseline methods in the Madmom library are pre-trained models, so there are no data requirements. For developing new methods, or adapting existing methods to TROMPA, we will use a number of public or private datasets for training and evaluation. These datasets are either public available datasets for scientific research such as the SMC Mirex Dataset [7] U. Gent Dataset [35] or MDW Piano corpus as well as data related to TROMPA such a the CDR catalog or synthetic datasets of choir singing related to **Deliverable D3.3 - Audio Processing**.

Human Annotations / Human Data

As mentioned above, in order to improve our methods we may ask humans (to be decided in each use case definition) to provide annotations using the Annotator Tool as:

- Correct the output of a beat tracking algorithm
- Verify if the output of the beat tracking algorithm is correct or not.
- Ask the to tap along a music piece (capture beats or tempo)
- Ask a user to annotate the time signature of piece or to verify that the output of an algorithm is correct (requires music experts)

Impact

The rhythm descriptors can be potentially use in all cases that involve audio files such as:

- **Music Enthusiasts**: Rhythm descriptors can be used to facilitate a music recommendation/music similarity engine and to search for music based on rhythm properties (e.g. tempo, time signature).
- **Instrument Players**: Rhythm descriptors such tempo fluctuations can be used to analyze existing or new performances of instrument players.
- **Choir Singers:** As in the instrument players use case, we can use rhythm descriptors to analyze rhythm characteristics of choir singing performances.

However the main contribution of this subtask is that we will run the rhythm analysis algorithms on a large portion of TROMPA data that is referred to the CE in the context of enrichment and create new reusable data for future scientific/musicological research.

2.2.4 Singing Voice Analysis

Summary

Singing voice analysis is the process of extracting information about the vocals of a music piece. These vocals may appear in diverse forms: accompanied by other instruments or a cappella, solo or in an ensemble, e.g., choirs, small vocal ensembles. Several expressive properties of singing voice can be extracted, including (but not limited to):

- **F0 contours** are time series showing the evolution of the fundamental frequency (F0) in time, usually expressed in Hertz (Hz). We obtain these contours using mono/poly-phonic F0 estimation algorithms. We also use F0 curves and F0 trajectories to refer to the same element.
- **Intonation**: measures the accuracy of pitch in singing concerning a specific tuning system, e.g., equal temperament or just intonation.

For ensemble singing, we can extract a set of additional properties:

- **The degree of unison** [8] can be computed from a performance where all voice sources sing the same notes, and we can measure it as the agreement between the F0 each source produces.
- **Synchronization between singers**: using the F0-curves and intonation descriptors, as well as other information such as note boundaries from scores, we can measure the synchronization between singers using various measures such as correlation.
- **F0 estimation and validation** in unison performances: employing an F0 estimation algorithm, we can extract the F0 curve of a unison, and study how it compares to the F0 curve from each singer individually.

Method Descriptions

We use a combination of existing and novel methods to tackle each of the tasks described in the previous section. For the extraction of **F0 contours** in monophonic signals, i.e., individual singers, we use state-of-the-art methods such as pYIN, [36], SAC [10], and CREPE [9]. In the case of **multiple F0 estimation** in choral music, we are currently developing a convolutional neural network (CNN) trained mainly on vocal quartets from various datasets (see the next section for a description).

There are several approaches to compute **intonation** features from a singing voice signal. It is often calculated frame-wise as the difference between the produced F0 - extracted from the audio signal - and the target pitch defined in the score of the piece [37].

In previous studies, we modeled the **synchronization** between singers using a linear correlation coefficient between the derivatives of the F0 contours [14], and using the mutual information measure between the deviation of the F0 contours from the score [37]. Additionally, we conducted a case study on the degree of unison in [14], where we extracted statistics from the distribution of F0 values at each frame.

To investigate **unison performances** further, we are currently studying how monophonic F0 estimation systems behave in such cases where several vocal sources produce the same notes. Using the individual F0 curves (from the datasets ground truth annotations), we compare the output of the F0 estimation algorithm to the F0 curve of each singer and the average.

Adaptation for TROMPA

In the context of singing voice analysis, these are the central adaptations we are doing for TROMPA:

- For multiple F0 estimation, we base our convolutional neural network on DeepSalience [11]. However, we introduce a few changes: we add the phase differentials as input features, we apply some modifications to the network design, and most importantly, we train the CNN on polyphonic vocal music from various datasets.
- For singer synchronization, we exploit intonation features and adapt some of the methodologies described in [12] and [13]. In these papers, the authors investigate several ways of measuring the synchronization between musicians of a string quartet; in the scope of ensemble singing, we adapted some of these methods and applied them to vocal music [14, 37].

Contribution to the State-of-the-Art

In the context of singing voice analysis, we mainly contribute to the state-of-the-art in multiple F0 estimation in polyphonic vocal music. The first version of this model will be capable of extracting one F0 value for each voice section in an audio recording, which is the first step towards a system for automatic transcription of choir recordings.

Additionally, and taking advantage of the human-annotated data that will be gathered in the scope of TROMPA, we plan to contribute to the state-of-the-art of singing performance rating, combining low-level and pitch descriptors with annotations.

In terms of data, besides the Choral Singing Dataset [14], two additional datasets are related to TROMPA:

- ESMUC Choir Recordings. Set of recordings of the ESMUC choir singers done in the scope of the project. The publication supporting these recordings is not yet available; however, we are actively working on curating these data to have a release at the end of this year.
- Dagstuhl ChoirSet. Singing voice dataset and accompanying paper submitted to a journal. This project is developed in collaboration with researchers from the International Audio Laboratories Erlangen (Germany). This paper is currently under review.

Data requirements

We exploit all the datasets described in the previous sections for various tasks. Given that all of them contain ground truth annotations for F0 contours, polyphonic audio mixtures, and stems, they are suitable for all tasks related to F0: (multiple) F0 estimation, intonation analysis, and unison description.

Human Annotations / Human Data

We plan to build a method to rate performances by a singer based on the singing analysis descriptors we described. Although this work will be based on the outcomes of Task 3.2, it is actually a part of Task 5.4 - Music Performance Assessment of WP5 and these outcomes will be reported on the corresponding deliverables. For the development of these methods, we may gather annotations about the quality of singing recordings. This human-labeled data will be combined with audio descriptors to automate this process. We will need the annotators to be music experts.

Impact

The singing voice analysis task is primarily related to the Choir Singers use case. However, some descriptors such as F0 or intonation, might also be used in other use cases, e.g., instrument players. Apart from the use of this data to the use cases, datasets created in the scope of this subtask will be publicly available, facilitating further research on the topic. Furthermore, we expect the intonation analysis of singing performances to be helpful in the context of individual choir rehearsals, serving as feedback for the singer.

2.2.5 Emotion Tag Annotation

Summary

General emotion recognition models may achieve a certain accuracy level, but there is need to further study agreement in subjective emotion annotation on data sets, with respect to musical preference, style and personal characteristics of the users. Two general approaches have been used in the conceptualization of emotions in music [29]:

- Categorical approach: This approach considers that there are a limited number of emotion categories, from which other emotions can be derived [30]. Major drawbacks of this approach are that the number of primary emotions categories results too small compared to the richness of music emotion perceived by humans and the high ambiguity of using language to describe human emotions.
- Dimensional approach: This approach considers that emotion can be modeled in two dimensions: valence (pleasantness or positiveness) and arousal (energy or activation) [31]. This approach can also be separated into four distinct quadrants: Q1 - positive valence and arousal, Q2 - negative valence and positive arousal, Q3 - negative valence and arousal, and Q4 - positive valence and negative arousal. Major drawbacks to this approach are the low agreement of the annotation of emotions from a musical excerpt and the blurriness of important psychological distinctions (such as anger and fear).

Recent research [18] has aimed to analyze agreement using a categorical approach, which allows less granularity than the dimensional approach, but can be used to analyze the subjective agreement amongst annotators. In TROMPA we plan to analyze agreement of annotations with respect to musical style, preference, and cultural background, and to widen this research to be more representative of the music enthusiasts audience and the TROMPA repertoire.

Method Description

We designed a Sparse Convolutional Autoencoder (SCAE) with Rectified Linear Unit Activations (ReLU). The dimensionality of an input mel-spectrogram feature (1 x 128 x 31) is increased to (128 x 2 x 31) in the latent space, by three double conv-layers augmenting the number of filters in the encoder: 32, 64, and 128, respectively. Dropout is set to 0.25 after every double conv-layer to prevent overfitting. Additionally, max-pooling and up-sampling are used to diminish and augment the dimensionality of the features with a variable pool size. Batch normalization is applied after each non-linearity to address internal covariate shift during training. The classifier is implemented by adding a flattening layer, 3 fully connected layers each with 512 neurons, followed by a Dropout layer each. Since we perform multi-task learning (MTL), we add three blocks of 2 fully connected layers (512) followed by a Dropout layer each, and three output layers with softmax activation. We implement MTL, since optimizing losses in the auxiliary tasks, can help improve generalization upon a main task. Each block represents a classifier: (1) quadrant prediction (4 classes, one per quadrant),

(2) arousal prediction (positive: Q1 and Q2, negative: Q3 and Q4), and (3) valence prediction (positive: Q1 and Q4, negative: Q2 and Q3). Following the FAIR disciplines of TROMPA we made the trained models available for testing online[4].

Adaptation for TROMPA

The Music Enthusiasts pilot has been designed to incentivise participation of users through the explanation of musical properties that relate to emotions. In the context of the TROMPA project, we will make use of the categorical approach to annotate musical excerpts with emotion ratings. We will study the agreement in emotion annotation and considering personalized models. Since we aim at using public domain music, we focus on the annotation of choral music. Through the use of the Contributor Environment, we will be able to connect to the Muziekweb website and retrieve excerpts of this type of music for annotation. In this context, our approach is twofold:

- Analyze the agreement of annotations by the music enthusiasts that participate in our application.
- Use these annotations to improve the performance of emotion recognition algorithms through personalization methods, i.e. with active learning.

Contribution to the State-of-the-Art

Schedl et al. [18] have created up to 267 annotations of musical excerpts from Beethoven's Eroica and analyzed correlations between perceived emotions and different demographics, musical expertise, familiarity with the music, personality traits (Five Factor Model), and audio features. To expand on this work, the proposal is to analyze the agreement of annotations with respect to musical style, preference, and cultural background. The contributions of this research relates directly to our approach:

- When achieving a higher amount of annotations, the quality of the emotion ratings should improve. The obtained annotations will be made available publicly for research on emotion recognition on choral music.
- The sparse convolutional autoencoder that was proposed uses multi-task learning, which is a novel approach to improve deep learning algorithms. With improved annotations, we plan to use crowdsourcing and group-wise annotations to improve the performance of these models.
- Since our approach relies on data annotation, the use case has been proposed in such a way that users may learn informally about musical properties that relate to emotion. We consider this a contribution from music annotation initiatives for educational purposes.

Data requirements

As mentioned previously, we make a pre-selection from choral music from the Muziekweb repository. We aim at recommending new music to the users starting from rock genres of the same emotional content as the annotations.

Human Annotations / Human Data

In order to analyze annotation agreement of emotional tags, we will ask the user to provide annotation in the following tasks:

[4] https://github.com/juansgomez87/quad-pred

- Emotion adjectives: Transcendence, Peacefulness, Power, Joyful activation, Tension, Sadness, Anger, Disgust, Fear, Surprise, Tenderness.
- Emotion classes: positive and negative arousal/valence.

In order to characterize personal characteristics, we will consider asking the user to provide personal information through a survey such as:

- Demographics
- Mother language
- Musical expertise
- Musical preference
- Personality traits

Impact

This technology is directly related to the Music Enthusiasts use case which is centered in using emotional content with the educational purpose of explaining musical properties that relate to the expression of particular emotions. Apart from that, we will contribute to the scientific community with data that facilitates research in the emotion and music field. Usually emotion annotations are made by up to 5-10 annotators per excerpt. Our approach involves obtaining more annotations than usual of each excerpt in order to improve the quality of annotations, albeit reducing the amount of categories. By augmenting the number of annotations, we attempt to have better insight on emotion perception annotation and agreement across different populations of users.

2.2.6 Music Similarity and Music Recommendation

Summary

Music discovery in today's digital environments is highly influenced by information filtering technologies, such as search engines and recommender systems (RS)[40]. Emerged as tool for helping users in finding relevant content, and for contrasting the so-called information overload [41], nowadays recommender systems are pivotal in most of the music streaming platforms (e.g. Spotify, YouTube, Pandora). Both from academia and from industry, great interest has been posed on improving recommendation techniques for maximizing the music listeners' engagement during their listening experience [42].

The literature on music similarity (see previous deliverable TR-D3.2.) has been used by MIR practitioners for designing, building and deploying music RS. Indeed, content-based and context-based are two of the three main standard frameworks used in RS scenarios, together with collaborative-filtering (CF) techniques [43]. CF methods analyze user-item interactions for providing similar items' recommendation to similar users, a method initially proposed as "social information filtering"[44]. Due to the possibility to adapt CF algorithms in different contexts, they are often used in commercial applications also outside the music domain.

Focusing on the music domain, three main future directions in music recommendation research have been identified by the MIR community: 1) *Psychologically-inspired* music recommendation 2) *Situation-aware* music recommendation 3) *Culture-aware* music recommendation [42]. In the context of the ME use-case (see deliverable D6.7-1), we focus on the development of emotion-based recommendations, part of the direction 1), as presented in the next sections.

Method Descriptions

Music recommendations provided in the context of the ME use-case are based on an hybrid system which makes use of both content- (e.g. audio), context- (e.g. tags) and collaborative filtering methods. In particular, emotion information for providing recommendations can be exploited in different ways:

- Item-item: comparing the content of different tracks for establishing a degree of similarity among tracks in the repertoire considered (e.g. comparing features' distribution or embedding this information in a low-dimensional space) .
- User-item: associating the content of a track with users' annotation, and analyzing the correlations among track' musical features and users' annotated emotion.
- User-user: understanding how individuals associate emotions to a track in comparison with annotations provided by different users.

Recommeder systems can make use of this different kinds of information for providing listening suggestions to the user base, optimizing for a range of objectives, such as accuracy with regards to users' preferences, or diversity of the recommender outcomes.

Adaptation for TROMPA

The direction of the music RS implemented for TROMPA is twofold: on one hand for providing recommendation we exploit the information extracted by means of automatic emotion-recognition algorithms. On the other hand, we compute community-based recommendations analyzing the crowdsourced ME participants' emotion annotations (see Section 2.2.6). In particular, emotion-related information is used as following described:

- At an item-level, tracks are associated with both content descriptors indicating emotion dimensional values (i.e. arousal and valence, see Section 2.2.6), and contextual tags (provided by users annotation, and external publicly available knowledge-bases).
- At a user-level, annotators' choices expressing emotion agreement are aggregated for characterizing individuals preferences and attitudes by means of comparing it with the community trends.

Recommendations are designed for giving users' new references for comprehending what are the facets that characterize emotion associated with a musical piece, providing a tool for expanding their awareness of the complexities of categorizing emotion when referring to music.

Contribution to the State-of-the-Art

By comparing the recommendations generated by the algorithmic-framework for emotion recognition and the community-based framework, we are able to identify what are limits and benefits of the frameworks considered. Moreover, joining the information provided by the two frameworks, hybrid models can advance state-of-the-art recommendation research, evidencing the tradeoff between the needs for automatic procedures and human annotation while studying the emotion associated with music.

Data requirements

We use the following data for computing music recommendations:

- Audio and metadata

- ➢ CDR music catalogue
- ➢ External publicly available knowledge-bases (e.g. MusicBrainz, last.fm)

❖ Emotion information

- ➢ Automatically extracted from the audio content
- ➢ Annotated by ME use-case participants

Impact

The recommendation framework implemented is tailored for the ME use-case, where its role is to help participants in: 1) discover new music; 2) understand how emotion is associated with music content. The interactions between ME use-case participants and provided recommendation can be used for advancing state-of-the-art music recommendation. Furthermore, anonymized users' explicit feedback data (likes, ratings, etc.) will be a resource for advancing user modelling and profiling for further music recommendation research.

2.3 Symbolic Descriptors

Summary

Symbolic descriptors will be used in facilitating tasks related to the **automatic assessment of music pieces** for the choir singers and the instrument players use cases. Moreover we extract baseline symbolic descriptors (onset density, interval histograms etc). These features will be used to the estimation of the difficulty of a music piece, and we will investigate the potential use to other use cases such as the music scholars use case.

Method Descriptions

The symbolic descriptors we use are based on conventional MIDI processing libraries such as MIDI Toolbox [19] and Pretty-Midi[5]. Apart the explicit use of these features for the use cases More precisely, our methodology includes

- ❖ **Rhythm analysis from symbolic streams**: We develop a method for automatic extracting rhythm information (beats, tempo) from symbolic MIDI streams [Ref Dafx].
- ❖ **Difficulty assessment of choir pieces**: By using Machine Learning methods, we will research which of the baseline features are related to the difficulty of a choir singing music piece based on its score.
- ❖ **Difficulty assessment of piano pieces**: Similar to the choir pieces, but for piano.

Although the difficult assessment tasks are related to Task 5.4 Performance Assessment, these will be based on the outcomes of this subtask of Task 3.2.

Adaptation for TROMPA

There is no special adaptation of the symbolic features to TROMPA. We only have to mention that the software used for the extraction of these features will be integrated into TROMPA Processing Library (Deliverable 5.3).

[5] https://github.com/craffel/pretty-midi

Contribution to the State-of-the-Art

We will contribute to new methods of processing MIDI streams for rhythm analysis.

Data requirements and Human Annotations / Human Data

We consider any type of symbolic data files (MIDI, MusicXML, MEI) of the pieces to analyze for analysis. These data can be enriched with additional information derived from TROMPA users such as:

- **Annotations** about the **difficulty** of the pieces, or specific sections of a music made by either music experts/conductors or amayeur singers/players.This would consist of labelling segments of a piece which are difficult.
- **Rhythm Annotations:** These annotations (such as beats, tempo, tempo fluctuations) can either extracted:
  - Automatically from a MIDI file. For example derive beat locations directly from MIDI (in case of aligned MIDI)
  - Use alignment methods and extract such information.
  - Annotations derived from humans.

Impact

The symbolic descriptors are initially focused on the choir singers and instrument players pilots. However they can be potentially used in other pilots, such as the music scholars pilots. This possibility will be investigated in the next months of the project and will be reported in detail in the next version of this deliverable.

2.4 Video Descriptors

Summary

Video descriptors are aimed to provide additional semantic information related to video recordings of musical performances. Often, video information can help to solve classical MIR tasks more accurately or without explicit supervision. In the scope of the project, we focus on the following tasks revealing the potential of using video data:

- **Video tagging**: general-purpose video tagging in the domain of music performances, providing frame-level and video-level labels from a pre-defined ontology.
- **Musical instrument detection:** providing a position of an object in the form of a rectangular bounding box.

Method Descriptions

Our method for video tagging and musical instrument detection uses Convolutional Neural Networks (CNNs) applied to log-mel spectrograms and video frames. In particular, we developed a musical instrument classification method [22] which is based on a multi-modal late fusion of audio and video features. The data is processed with two individual networks (Xception-like architecture for audio analysis and VGG-like architecture for frames analysis) and the learned features are used in an additional fully-connected layer to obtain the classification results. Furthermore, in order to facilitate interpretability of the results obtained in [22], we conduct a correlation analysis in [38], [39] which

couple the network predictions with the regions where the instrument appears (in case of video frames) or with a characteristics patterns in the spectrograms (in case of audio data).

Adaptation for TROMPA

The developed methods are aimed for the Music Enthusiasts Use-Case. The direction of the music RS implemented for TROMPA is twofold: on one hand for providing recommendation we exploit emotion information extracted by means of automatic emotion-recognition algorithms. On the other hand, we compute community-based recommendations analyzing the crowdsourced ME participants' emotion annotations (see Section 2.2.6).

Contribution to the State-of-the-Art

In the context of TROMPA, we mainly contribute with the multi-modal CNN-based algorithm for multi-label video tagging [22] and the interpretability methods [38], [39] which are suitable for the Music Enthusiasts Use-Case. In this case, we can provide additional context information of the instruments which are present in the recording, highlight and label them, therefore increasing the listener's involvement and facilitate the possibility of following individual musical tracks.

Data requirements

Depending on the task, we may need to adapt existing datasets or gather new ones. For video tagging, we can use general-purpose datasets of video recordings (or their subsets containing musical performances) such as AudioSet [25], Youtube-8M [26], FCVID [27] as well as high-quality music-related datasets for the evaluation such as URMP [28].

Human Annotations / Human Data

Human annotations and evaluation may be needed for the following tasks (to be defined upon the selection of the use cases):

- Collecting datasets (providing bounding boxes, labels, masks)
- Providing an upper boundary of quantitative evaluation (in a form of inter-annotation agreement rate)
- Subjective (qualitative) evaluation of the developed methods

Data output format

Descriptor	Representation
tags	One text file per video file, containing pairs (timestamp, tags) where tags are a set of labels from a pre-defined ontology
detection	One text file per video file, containing pairs (timestamp, bbox) where bbox is represented as a set of coordinates (left, bottom, right, top)
segmentation	K binary masks per timestamp where K is a number of detected objects in a frame at the timestamp

Impact

Video descriptors can be used in many use cases which involve video, to name a few:

- **Instrument Players**: showing the fingering charts upon a video recording.
- **Choir Singers**: counting the number of singers.
- **Music Enthusiasts**: providing a more detailed transcription for music performances, instrument-to-instrument navigation in music videos (should be paired with source localization/separation/re-mixing), and highlighting playing/non-playing instruments in videos.

3. Automatic Description Workflow

In this section we will provide some information about the automatic description workflow, how the CE will interact with the various music description tools discussed in the previous section, and how the music description tools will store the results. This is a part of **Deliverable 5.3 - TROMPA Processing Library**, which describes how the music description technologies are integrated to the CE. In this section we will briefly describe the process for reasons of clarity.

Each descriptor extraction program is represented in the CE database as a set of nodes. These nodes are related to items such as the software that implements the algorithm, the algorithm (a software can contain more than one algorithm) or the parameters of it. Each descriptor extraction program consists of:

- A **SoftwareApplication** node that represents the software/library that hosts the specific algorithm.
- An **EntryPoint** node that corresponds to a specific algorithm/method to be run within the SoftwareApplication.
- An *actionApplication* relation between **SoftwareApplication** and **EntryPoint** that defines that this EntryPoint is a part of the **SoftwareApplication** node**.**
- A **ControlAction** template node that is the template of the description task to be executed. Each time an algorithm is run, a copy of this template node is created (instantiated). It can be viewed as the *interface* of the algorithm to be run.
- A *potentialAction* relation between **EntryPoint** and **ControlAction.**
- **Property** nodes that correspond to the items (e.g. audio files) in the CE database that are to be processed.
- **PropertyValueSpecification** nodes which correspond to a scalar parameters (string, number, on/off checkbox) that need to be given by the user as 'settings' inputs, in order to tune the algorithm process.
- Relations between **ControlAction** and **Property**/**PropertyValueSpecification** template nodes.

From the algorithm/software side, each algorithm has to:

- Create a websocket connection to the CE and use GraphQL's *subscription* functionality to register itself as an algorithm for a given EntryPoint. This functionality offers the possibility to handle user requests for algorithms in real-time.
- Frequently check for tasks in the CE by querying for **ControlAction**s created on the basis of a certain **EntryPoint**.
- Once the algorithm receives the job, it can update the status of the ControlAction node to 'received'. While the algorithm is running, it can update the status of the ControlAction node in order to provide more information (e.g. 'running') . Once the process has completed, the

algorithm process application should write the result to a public location and add a reference to this result in the CE database:

- When the algorithm process application now updates the **ControlAction** actionStatus to 'complete', the process request response cycle is completed.

Apart from the representation of the algorithms on the CE, for each algorithm the following has to be decided:

- **Software Specifications**: Specify the software, the OS to be run and other software dependencies.
- **Hardware Specifications:** Specify hardware requirements (e.g. GPUs for Deep Learning based algorithms, number of CPU cores etc)
- **Physical Location:** Where it is executed. Decide on which computer(s) the algorithm should be executed. We can have a single machine running multiple algorithms in parallel, or many computers running specific algorithms. This flexibility is ensured by the design principles of the **TROMPA Processing Library**. In any case the hardware requirements should be satisfied.
- **Parameters:** Specify the parameters of the algorithm (we can have multiple instances of the same algorithm with different parameters) and the arguments used (input, output files).
- **Descriptor Dependencies:** Specify if it requires other descriptors: the extraction of one descriptor can be based on other descriptors (sequential processing). For example to extract Pitch Class Profiles a chroma-based spectrogram is needed.
- **Storage:** Where the descriptions are stored. There are two options:
  - **Internal:** Store the descriptors as a node in the CE. This is more appropriate for more high level descriptors (e.g. tempo, time signature, emotion tags).
  - **External**: Store descriptors to external data repositories. This is appropriate for low-mid level descriptors. For example storing frame level features (e.g. mfcc) as nodes in the CE is not efficient in terms of space and query time execution.
  - **Hybrid**: Store binary values (as in the case of frame-level features) as textual information in a CE node.

These specifications are related to the integration of the algorithms rather than their design and thus will be defined by the end of the project and will be reported in the 2nd version of **Deliverable D5.3 - TROMPA Processing Library.**

4. Conclusion

This deliverable described the technologies used for the automatic description of TROMPA music material. The main source of information is the **audio** itself, but we are not limited to this and we also consider **symbolic** and **video** descriptors that are used to facilitate the use cases. Moreover, apart from the use cases, these descriptors will be used to enrich the TROMPA content for future research and future potential use cases. By the end of the project these descriptors will be deposited in various repositories linked with the Contributor Environment as contributions of TROMPA to public domain musical archives (e.g. automatic annotations, low-level descriptors).

Additionally, all algorithms will be "packaged" as components of the **TROMPA Processing Library**. This approach has two major advantages:

- Use cases will be "agnostic" of where and how the algorithms are run and can only make requests to the CE to trigger the appropriate algorithms and receive the descriptors needed.

- ❖ Algorithms will be potentially available to be re-used by future pilots/applications after the end of the project.

Although this is the final version of this deliverable, scientific research on the topics described in this deliverable will be continued until the end of the project. Any further scientific contributions such as scientific publications or datasets will be published on the corresponding webpage[6] of our website and reported on the periodic reports.

5. References

5.1 Written references

[1] Bogdanov, D., Wack N., Gómez E., Gulati S., Herrera P., Mayor O., et al. (2013). ESSENTIA: an Audio Analysis Library for Music Information Retrieval. International Society for Music Information Retrieval Conference (ISMIR'13). 493-498.

[2] Böck, Sebastian, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. "Madmom: A new Python audio and music signal processing library." In *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 1174-1178. ACM, 2016.

[3] Böck, Sebastian, and Markus Schedl. "Enhanced beat tracking with context-aware neural networks." In *Proc. Int. Conf. Digital Audio Effects*, pp. 135-139. 2011.

[4] Gkiokas, Aggelos, and Vassilios Katsouros. "Convolutional Neural Networks for Real-Time Beat Tracking: A Dancing Robot Application." In *ISMIR*, pp. 286-293. 2017.

[5] Gouyon, Fabien, Anssi Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano. "An experimental comparison of audio tempo induction algorithms." *IEEE Transactions on Audio, Speech, and Language Processing* 14, no. 5 (2006): 1832-1844.

[6] Tzanetakis, George, and Perry Cook. "Musical genre classification of audio signals." *IEEE Transactions on speech and audio processing* 10, no. 5 (2002): 293-302.

[7] Holzapfel, Andre, Matthew EP Davies, José R. Zapata, João Lobato Oliveira, and Fabien Gouyon. "Selective sampling for beat tracking evaluation." *IEEE Transactions on Audio, Speech, and Language Processing* 20, no. 9 (2012): 2539-2548.

[8] Sundberg, Johan, and Thomas D. Rossing. "The science of singing voice." (1990): 462-463.

[9] Kim, Jong Wook, Justin Salamon, Peter Li, and Juan Pablo Bello. "CREPE: A Convolutional Representation for Pitch Estimation." *arXiv preprint arXiv:1802.06182* (2018).

[10] Villavicencio, Fernando, Jordi Bonada, Junichi Yamagish, and Michel Pucher. *Efficient Pitch Estimation on Natural Opera-Singing by a Spectral Correlation based Strategy*. Technical report, 2013.

[11] Bittner, Rachel M., Brian McFee, Justin Salamon, Peter Li, and Juan Pablo Bello. "Deep Salience Representations for F0 Estimation in Polyphonic Music." In *ISMIR*, pp. 63-70. 2017.

[12] Papiotis, Panos, Marco Marchini, and Esteban Maestre Gómez. "Multidimensional analysis of interdependence in a string quartet." In *Williamon A, Goebl W, editors. International Symposium on Performance Science (ISPS); 2013 Aug 28-31; Vienna, Austria. Brussels: Associoation Européenne des Conservatoires; 2013. p. 563-8.* Associoation Européenne des Conservatoires, 2013.

[6] https://trompamusic.eu/research/results/publications

[13] Papiotis, Panos, Marco Marchini, Alfonso Perez-Carrillo, and Esteban Maestre. "Measuring ensemble interdependence in a string quartet through analysis of multidimensional performance data." *Frontiers in psychology* 5 (2014): 963.

[14] Cuesta, Helena, Emilia Gómez Gutiérrez, Agustín Martorell Domínguez, and Felipe Loáiciga. "Analysis of intonation in unison choir singing." (2018).

[15] Bertin-Mahieux, Thierry, Daniel PW Ellis, Brian Whitman, and Paul Lamere. "The Million Song Dataset." In *Ismir*, vol. 2, no. 9, p. 10. 2011.

[16] Porter, Alastair, Dmitry Bogdanov, Robert Kaye, Roman Tsukanov, and Xavier Serra. "Acousticbrainz: a community platform for gathering music information obtained from audio." In *International Society for Music Information Retrieval Conference*. 2015.

[17] Vigliensoni, Gabriel, and Ichiro Fujinaga. "The music listening histories dataset." In Proceedings of the 18th International Society for Music Information Retrieval Conference. Suzhou, People's Republic of China, 2017

[18] Schedl, Markus, Emilia Gómez, Erika S. Trent, Marko Tkalčič, Hamid Eghbal-Zadeh, and Agustin Martorell. "On the Interrelation between listener characteristics and the perception of emotions in classical orchestra music." *IEEE Transactions on Affective Computing* 9, no. 4 (2018): 507-525.

[19] Eerola, Tuomas, and Petri Toiviainen. "MIDI toolbox: MATLAB tools for music research." (2004).

[20] Schedl, Markus, Arthur Flexer, and Julián Urbano. "The neglected user in music information retrieval research." *Journal of Intelligent Information Systems* 41, no. 3 (2013): 523-539.

[21] Volk, Anja, Elaine Chew, Elizabeth Hellmuth Margulis, and Christina Anagnostopoulou. "Music similarity: concepts, cognition and computation." (2016): 207-209.

[22] Slizovskaia, Olga, Emilia Gómez, and Gloria Haro. "Musical instrument recognition in user-generated videos using a multimodal convolutional neural network architecture." In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pp. 226-232. ACM, 2017.

[23] Senocak, Arda, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. "Learning to localize sound source in visual scenes." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4358-4366. 2018.

[24] Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., & Torralba, A. (2018). The sound of pixels. arXiv preprint arXiv:1804.03160.

[25] Gemmeke, Jort F., Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. "Audio set: An ontology and human-labeled dataset for audio events." In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 776-780. IEEE, 2017.

[26] Abu-El-Haija, Sami, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. "Youtube-8m: A large-scale video classification benchmark." *arXiv preprint arXiv:1609.08675* (2016).

[27] Jiang, Yu-Gang. "Categorizing big video data on the web: Challenges and opportunities." In *Multimedia Big Data (BigMM), 2015 IEEE International Conference on*, pp. 13-15. IEEE, 2015.

[28] Li, Bochen, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. "Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications." *IEEE Transactions on Multimedia* 21, no. 2 (2019): 522-535.

[29] Vuoskoski, J.K., and Eerola, T. “The role of mood and personality in the perception of emotions represented by music.” *Cortex*. 47(9), 1099-1106, 2011.

[30] Hevner, K. “Experimental studies of the elements of expression in music.” *American Journal of Psychology*, 48, 246-268, 1936.

[31] Russell, J.A. “A circumplex model of affect.” *Journal of Personal and Social Psychology*, 39(6), 1161-1178, 1980.

[32] Casey, Michael A., Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. "Content-based music information retrieval: Current directions and future challenges." *Proceedings of the IEEE* 96, no. 4 (2008): 668-696.

[33] Knees, Peter, and Markus Schedl. "A survey of music similarity and recommendation from music context data." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 10, no. 1 (2013): 2.

[34] Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 801-818. 2018.

[35] Varewyck, Matthias, and Jean-Pierre Martens. "Assessment of State-of-the-Art Meter Analysis Systems with an Extended Meter Description Model." In *ISMIR*, pp. 311-314. 2007.

[36] Mauch, Matthias, and Simon Dixon. "pYIN: A fundamental frequency estimator using probabilistic threshold distributions." 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.

[37] Cuesta, Helena and Gómez, Emilia. “Measuring Interdependence in Unison Choir Singing” in *Late-breaking Demo Session. ISMIR 2018.* Paris (France).

[38] Slizovskaia, Olga, Emilia Gómez, and Gloria Haro. "A Case Study of Deep-Learned Activations via Hand-Crafted Audio Features." The 2018 Joint Workshop on Machine Learning for Music, ICML 2018. Stockholm (Sweden).

[39] Slizovskaia, Olga, Emilia Gómez, and Gloria Haro. “Correspondence between audio and visual deep models for musical instrument detection in video recordings.” Paper presented at: 18th International Society for Music Information Retrieval Conference (ISMIR17); Late Breaking Demo, 2017 Oct 23-27; Suzhou, China.

[40] Schedl, M., Gómez, E., & Urbano, J. (2014). Music information retrieval: Recent developments and applications. Foundations and Trends in Information Retrieval, 8(2–3), 127–261.

[41] Bollen, D., Knijnenburg, B. P., & Graus, M. (2010). Understanding Choice Overload in Recommender Systems Categories and Subject Descriptors. Proceedings of the Fourth ACM Conference on Recommender Systems - RecSys ’10, 63–70.

[42] Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., & Elahi, M. (2018). Current Challenges and Visions in Music Recommender Systems Research. International Journal of Multimedia Information Retrieval, 7(2), 95–116

[43] Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. Knowledge-Based Systems, 46, 109–132.

[44] Shardanand, U., & Maes, P. (1995). Social information filtering: algorithms for automating “word of mouth.” Proceedings of the Conference on Human Factors in Computing Systems, 210–217.

5.2 List of abbreviations

Abbreviation	Description
HPCP	Harmonic Pitch Class Profile
ZCR	Zero Crossing Rate
BPM	Beats per Minute
RNN	Recurrent Neural Networks
LSTM	Long Short Term Memory
CNN	Convolutional Neural Networks
MIR	Music Information Retrieval
DL	Deep Learning
Partner	**Description**
UPF	University Pompeu Fabra
ESMUC	Escola Superior de Música de Catalunya
CDR	Centrale Discotheek Rotterdam