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 v1
Work Package	WP3 - Automated Music Data Processing and Linking
Deliverable Type	Report
Dissemination Level	PU- Public
Document due date	28 February 2019
Date of submission	28 February 2019
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 (UPF), Lorenzo Porcaro (UPF)
Reviewers	Vladimir Viro (PN)

Executive Summary

This aim of this task is to develop and integrate technologies for the automatic description of the majority of the musical data described and collected in **T3.1 Data Resource Preparation**. It's aim is to provide music descriptors at various modalities and levels of analysis for the target repertoires in order to facilitate the use cases. Although this task primarily deals with audio, we also consider symbolic and video music sources. The goals of this task can be summarized to:

- Apply and evaluate existing of the 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) and algorithms to open repositories such as AcousticBrainz and Zenodo in order to facilitate future research on music processing.

We will use various existing state-of-the-art methods/libraries for the extraction of descriptors. The aim of this deliverable is to provide descriptors for three type of music modalities, namely **audio**, **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 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 will be extracted and evaluated in the context of TROMPA use cases but we do not intend to further investigate and research the development of these descriptors. We will contribute to the SoA in terms of evaluating these descriptors described in large amounts of western classical music data. Moreover we may contribute to the SoA with new datasets (see next subsection Human Annotations / Human Data). The harmonic/tonal 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. We will use and adapt existing state-of-the-art methods of rhythm processing methods will be evaluated in the scope of TROMPA and if the adopted methods do not prove to be efficient, we will develop new methods. The rhythm descriptors 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.
- ❖ **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 will 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 plan to contribute to the state-of-the-art of singing performance rating, combining low-level and pitch descriptors with annotations.
- ❖ **Music similarity**: In the context of TROMPA project, we will make use of content and context-based models for retrieving similarity measures. In particular, we will focus on understanding the role of these measures when embedded in more complex architectures, such as Musical Recommender Systems. Indeed, we will make use of the notion of similarity for **understanding** and **characterizing** the complex and heterogeneous nature of the Western Classical Music repertoire, we will compare different musical repertoires, and we will create **listening experiences** for the users which can enhance the discovery of the European Musical heritage. Music similarity estimation can be potentially used in the following scenarios:
    - ➢ **Music Enthusiasts**: help the users in the discovery of Western Classical Music.
    - ➢ **Instrument Players, Choir singers**: help musician and singers during the learning process in identifying musical pieces which most can fit their education, basing on the training level and personal musical taste.

- **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 consider also personalized models. In order to analyze annotation agreement of emotional tags, we will ask the user to provide annotations in emotion ratings such as transcendence, peacefulness, power, joyful activation, tension, sadness, anger, disgust, fear, surprise, tenderness on a likert scale.
- **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. We will investigate which score features of a music piece are relevant to its playing difficulty. To do so, we will need several types of data to train and evaluate our model including **symbolic scores**, **annotations** about the **difficulty** of the pieces, and **audio recordings** of several renditions of the pieces performed by different people. 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.
- **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 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, highlighting playing/non-playing instruments in videos.

The automatic description tools that act on a file input (e.g. music file or video file) will be made available as executable programs or will be integrated to Essentia, allowing us to programmatically call the tools and retrieve the descriptors. We plan to develop a management tool that will allow us to automatically retrieve content described in the CE, perform some computation on that content, and then store descriptors or some other type of generated data, making it available again in the CE. For the tools that do not operate on a simple input/output basis we plan to provide documentation and software libraries for the developers of these tools to be able to easily retrieve data from the CE and submit descriptors or other results for other members of the consortium to use. We will promote the use of common file formats to store the descriptors computed by the tools presented in this document. A final decision for the data format will be described in the next version of this deliverable. We plan to publicly host the descriptors computed by these tools so that they can be accessed by other members of the consortium and the public. The location of these descriptors will be able to be obtained by querying the CE.

Version Log		
#	Date	Description
v0.1	14 February 2019	Initial version submitted for internal review
v0.2	21 February 2019	Revised version after internal review
v1.0	28 February 2019	Final version submitted to EC

Table of Contents

1. Introduction

This aim of this task is to develop and integrate technologies for the automatic description of the majority of the musical data described and collected in **T3.1 Data Resource Preparation**.

1.1 Scope

This deliverable is the first version of the Deliverable 3.2 - Music Description, part of Work Package 3. It’s aim is to provide music descriptors at various modalities and levels of analysis for the target repertoires in order to facilitate the use cases. 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**[1], and they are jointly submitted in M10 of the project. The final version of this deliverable will be submitted in M20.

1.2 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 of the 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) and algorithms to open repositories such as AcousticBrainz and Zenodo in order to facilitate future research on music processing.

2. Music Descriptors

2.1 Overview and Relation to the Use Cases

In this section we will describe the extraction of the music descriptors along with reference to existing open source repositories and publications. We will use various existing state-of-the-art methods/libraries for the extraction of descriptors. The aim of this deliverable is to provide descriptors for three type of music modalities, namely **audio**, **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.

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

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 Detection	Music Enthusiasts
	Tag 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.

In the next sections we will provide some details of the music description methods. In Section 2.2 we describe the **audio descriptors**. We will start from the **low level** audio descriptors, such a spectral and cepstral frame based descriptors. At next we will 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 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.

2.2 Audio Descriptors

Essentia [1] will be used as the as the main framework for extracting state-of-the-art 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 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.

2.2.1 Low Level 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 focus 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 will contribute to the SoA in terms of evaluating the harmonic/tonal descriptors described in large amounts of western classical music data. Moreover we may contribute to the SoA with new datasets (see next subsection Human Annotations / Human Data)

Data requirements

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

Human Annotations / Human Data

Although no new methods for Harmonic/Tonal description will be developed, during TROMPA use cases we may collect user data as:

- Correct the output of the chord progression system
- Evaluate the scale or the key of a music piece (correct or not)
- Ask users to annotate with chords or key signature segments of audio (requires experts).

These annotations 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

The harmonic/tonal 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 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

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)

Adaptation for TROMPA

We will use and adapt existing state-of-the-art methods of rhythm processing. More precisely, as a baseline we will use the Madmom [2] library, which provides various methods for rhythm tracking, mainly based on Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) networks. Madmom will be used for the tasks of beat tracking, tempo estimation and downbeat tracking. The proposed methods will be evaluated in the scope of TROMPA. If the adopted methods do not prove to be efficient, we will develop new methods based on LSTM [3] and Convolutional Neural Networks (CNN) [4],

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 the 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.

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 including:

- Beat Tracking:
    - Ballroom Dataset [5]
    - GTZAN Dataset [6]
    - SCM Mirex Dataset [7]
    - RWC Dataset[2]

- Tempo Estimation
    - IMSIR 2004 Songs Dataset [5]
    - Ballroom Dataset [5]
    - GTZAN Dataset [6]
- Time signature:
    - Ballroom Dataset [5]
    - RWC Dataset
    - U. Gent Dataset [35]

Human Annotations / Human Data

In order to improve our methods we may ask humans (to be decided in each use case definition) to provide annotations in the following tasks:

- 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)

Data output format

Table 2.3 summarizes the output format of the description methods that will be used to store the descriptors.

Descriptor	Representation
beats/downbeats	One text file per audio file, containing the position of the beats/downbeats
Tempo over time	One text file per audio file, containing time instants and the corresponding tempi (in BPM)
Tempo single	One text file per audio file, containing a single tempo value (in BPM)
Time signature	One text file per audio file, containing pairs of (time, time signature) for

[2] https://staff.aist.go.jp/m.goto/RWC-MDB/

	storing time signatures that may change in a single file.

Table 2.3. Rhythm descriptors representation.

Relation to the use cases

The rhythm 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**: 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.

2.2.4 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 (f0) in time, usually measured in Hz. Monophonic pitch estimation and multi-pitch estimation for transcription purposes.
- **Intonation**: Measures the accuracy of pitch in singing with respect to a specific tuning system, e.g. equal/non equal temperament or a specific reference pitch. This value is commonly computed frame-wise as the difference between the produced pitch and the target pitch defined in the score of the piece.

For ensemble singing, i.e. choral singing, additional specific properties can be extracted:

- **Degree of unison** or **pitch dispersion** in unison singing: The agreement between all the voice sources is the degree of unison [8], which is inversely proportional to the dispersion.
- **Synchronization** between singers: using the pitch and intonation descriptors, as well as other information such as the note boundaries that can be obtained from the score, the synchronization between singers can be quantified using correlation-like measures.

Adaptation for TROMPA

We will use and evaluate state-of-the-art algorithms for pitch and multi-pitch estimation, and adapt them to singing voice. Precisely, we will use CREPE [9] and SAC [10] for monophonic singing and Deep Salience [11] for polyphonic singing. These architectures will be evaluated in the context of TROMPA and if they do not prove to be efficient, we will adapt them using different training data. More specifically, Deep Salience will be re-trained with polyphonic singing voice data to make it more suitable for multi-pitch estimation in choral singing. Specific methodologies for singing intonation and measuring the synchronization between singers will be developed based on [12] and [13].

Contribution to the State-of-the-Art

We will contribute to the state-of-the-art in multi-pitch estimation, building new models trained using choral singing data, which is more challenging than other types of data such as piano or guitar music. Regarding the intonation analysis, 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 plan to contribute to the state-of-the-art of singing performance rating, combining low-level and pitch descriptors with annotations.

Data requirements

For the development of new methods we will use a number of datasets for training and evaluation:

- Choral singing dataset [14].
- ESMUC[3] Choir dataset. Recorded in the scope of TROMPA.

Both datasets contain individual tracks, choir mixes, and scores. This makes them suitable for the pitch and multi-pitch estimation, and also for the intonation description.

Human Annotations / Human Data

In order to build a method to rate a performance by a singer, 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.

Relation to the use cases

This particular task is primarily related to the Choir Singers use case. However, some descriptors such as pitch or intonation might also be used in other use cases, e.g. instrument players.

2.2.5 Music Similarity

Summary

The task of assigning a similarity degree between two or more musical items is an ill-defined problem, considering the high subjectivity of the notion of similarity when referring to music [20]. However, in the field of Music Information Retrieval (MIR) to model the notion of music similarity has become fundamental when facing problems such as content-based querying and retrieval, automatic music classification or music recommendation [21]. Within several factors which can influence the notion of similarity, we can identify four main categories, two music-related and two user-related:

- Music Content (rhythm, melody, armony, etc.)
- Music Context (song lyrics, artist’s background, semantic labels, etc.)
- User Properties (demographics, musical preferences, musical training, etc.)
- User Context (mood, social context, activities, etc.)

According to the MIR literature, methods for estimating music similarity can be divided in two families:

[3] Escola Superior de Música de Catalunya (ESMUC) is an associated partner of the TROMPA project

- Content-based similarity method: makes use of several descriptors (Low level,Harmonic/Tonal, Rhythm, see Sections 2.2.1-3) which can be extracted from the audio signals. In this case, the main advantage relies on the possibility to estimate the similarity using only the audio file. Moreover, subjective bias are not included, considering the objective nature of the acoustic features. However, these methods lack of semantic information intrinsically contained in the human cognition of similarity, the so-called problem of the "semantic-gap"[32].
- Context-based similarity methods: makes use of difference sources (Web pages, tags, playlists, metadata) which can captures similarity aspects beyond the audio signal. These methods are generally more aligned with user preferences, being partly derived from community-based sources or listening behaviours. In this case, the main disadvantages derive from the noisiness and the subjectivity of the data considered when building the models[33].

Adaptation for TROMPA

In the context of TROMPA project, we will make use of both content and context-based models for retrieving similarity measures. In particular, we are interested in understanding the role of these measures when embedded in more complex architectures, such as Musical Recommender Systems. Indeed, we will make use of the notion of similarity for:

- Understanding and characterizing the complex and heterogeneous nature of the Western Classical Music repertoire.
- Comparing different musical repertoires, both from a content perspective and user preferences, aiming to discover implicit and explicit links.
- Creating listening experiences for the users which can enhance the discovery of the European Musical heritage.

Contribution to the State-of-the-Art

Classical Music characterization and discovery is extremely influenced by the audio descriptors, in the content-based models, and user behaviours, in the context-based models. Consequently, music similarity inherits a series of issues which limits the effectiveness of its measuring, especially when embedded in more complex system. Using the data gathered within the TROMPA project, we aim to adapt and to improve the existing models for estimating the music similarity in the case of Western Classical Music.

Data requirements

Datasets needed for estimating music similarity are of two kinds:

- Audio files: both tracks from classical repertoire and from different repertoires will be used for extracting several descriptors. Among the others already identified:
  - Million Song Dataset [15]
  - Acousticbrainz [16]
  - Muziekweb[4]
- Users logs: user listening histories will be considered for understanding and analyzing preferences and tastes:

[4] https://www.muziekweb.nl/

- ➢ The Music Listening Histories Dataset [17]
- ➢ Listenbrainz[5]
- ➢ Muziekweb

In addition, we will make use of other sources publicly available for gathering more musical contextual information, as instance Wikipedia or MusicBrainz, and also Web API, such as Spotify and Last.fm.

Human Annotations / Human Data

We will ask humans for two kinds of contributions:

- ❖ Explicit feedback: users will be asked to assess the degree of similarity between two or more tracks
- ❖ Implicit feedback: we will analyze the user behaviours during the listening experiences, deriving measures of similarity.

Relation to the use cases

Music similarity estimation can be potentially used in the following use cases::

- ❖ Music Enthusiasts: it can be used for helping the users in the discovery of Western Classical Music.
- ❖ Instrument Players, Choir Singers: it can be used for helping musician and singers during the learning process in identifying musical pieces which most can fit their education, basing on the training level and personal musical taste.

2.2.6 Emotion Tag Annotation

Summary

General emotion recognition models may achieve 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) and arousal (activation) [31]. 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.

[5] https://listenbrainz.org/

Adaptation for TROMPA

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 also consider personalized models.

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.

Data requirements

We plan to run this experiment on the Muziekweb website, taking advantage of historical user logs to evaluate music preference from the annotators. Following Schedl et al., we propose to use particular 30 second excerpts, which should be selected with the evaluation from musicologists, such as for instance:

- Muziekweb Users logs: Both the lending history and Rating logs stored by CDR can be used to evaluate users musical preference with respect to classical music and other musical styles.
- Muziekweb Audio files: Excerpts of Mahler's 5th Symphony and Uri Caine's arrangement of the same music piece in different style to analyze the transfer of emotions within styles.

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:

- Emotion ratings: Transcendence, Peacefulness, Power, Joyful activation, Tension, Sadness, Anger, Disgust, Fear, Surprise, Tenderness on a Likert scale.

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

- Demographics
- Language
- Musical expertise
- Musical preference
- Personality traits

Relation to the use cases

Music emotion classification is aimed at the following scenario:

- Music Enthusiasts: it can be used for providing users information about the life and work of the selected composer, e.g. Mahler.

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 will extract baseline symbolic descriptors that can be potentially exploited in other use cases, such as the music scholars use case. Starting from existing MIDI processing libraries such as MIDI Toolbox [19] and Pretty-Midi[6] and we will investigate which score features of a music piece are relevant to its playing difficulty. More precisely, our methodology includes

- **Extraction of baseline symbolic features** (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.
- **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 its score.
- **Difficulty assessment of piano pieces**: Similar to the choir pieces, but for piano.

More details on the methods that will be developed will be provided in the next version of this deliverable.

Data requirements

We will need three types of data to train and evaluate our models:

- **Symbolic scores** (i.e. MIDI) of the pieces to analyze.
- **Annotations** about the **difficulty** of the pieces, ideally made by music experts/conductors.This would consist of labelling segments of a piece which are difficult.
- **Audio recordings** of several renditions of the pieces performed by different people.

Using the audio recordings and their associated synchronized scores, we plan to exploit the descriptors presented in Section 2.2.4 (Singing Voice Analysis) to identify potentially difficult parts of a piece as those where intonation and timing deviates the most from the score. This information, together with the one provided by difficulty annotations, will be used as ground truth data to build Machine Learning models to automatically assess the difficulty of a piece directly from the score.

Human Annotations / Human Data

During the **choir singers** and **instrument players** pilots, we will gather data regarding the difficulty of music pieces. These data will be a contribution of TROMPA to public domain archives, as well as will be used for training/evaluating the prediction models.

Relation to the use cases

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.

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

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 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.

Adaptation for TROMPA

For video tagging and instrument detection, we will use a method from [22] which is based on a multi-modal fusion of audio and video features. For localization, we can advance with modern unsupervised methods such as [23] or [24]. For general-purpose segmentation, it’s more reasonable to build a system upon existing Deep Learning (DL) solutions like DeepLabv3+ [34]. However, other classical computer vision methods and frameworks [OpenCV] can be used for more specific tasks e.g. skin detection.

Contribution to the State-of-the-Art

We can advance multi-label video tagging (with audio as an additional source). In the case of localization and segmentation, we could provide a first-ever SoTA for multimodal unsupervised methods but only by the mean of creating a new high-quality (not necessarily self-recorded) dataset for evaluation.

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].

For the task of object localization and segmentation, we would need to gather or record a new dataset for training and evaluation. Another possibility is to do a manual selection of a subset from one of the general-purpose datasets and annotating it.

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

Relation to use cases

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 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.

3.1 Communication with the Contributor Environment

Description tools that act on a file input (e.g. music file or video file) will be made available as executable programs or will be integrated to Essentia, allowing us to programmatically call the tools and retrieve the descriptors. We plan to develop a management tool that will allow us to automatically retrieve content described in the CE, perform some computation on that content, and then store descriptors or some other type of generated data, making it available again in the CE. This tool will be developed in the scope of **Deliverable 5.3 - TROMPA Processing Library**[7], which will be submitted as a deliverable in two versions, one on M12 and one on M34.

3.2 Storage of Descriptors

3.2.1 Data format

We will promote the use of common file formats to store the descriptors computed by the tools presented in this document. While we have not finalised the formats to be used by these tools, we expect to use open formats that are already commonly used in the Music Information Retrieval community and that are easily consumable by a wide range of software and software libraries. A final decision for the data format will be described in the next version of this deliverable.

3.2.2 Repositories

We plan to publicly host the descriptors computed by these tools so that they can be accessed by other members of the consortium and the public. While the final location is not finalised we expect that they will be hosted either on infrastructure provided by the task leader (UPF) or in the storage component provided by the Contributor Environment. The location of these descriptors will be able to be obtained by querying the CE.

When the result of a research task results in the improvement to an algorithm which is part of essentia, we will promote its inclusion in the essentia MusicExtractor[8], which will in the future be used in other projects such as AcousticBrainz, contributing to the quality of data in this repository.

When we create datasets of examples of audio for certain research tasks we will also release the contents of this dataset on Zenodo if the license of the content allows, which will let researchers from other institutions to take advantage of the resources that we created.

4. Conclusion

This deliverable describes the technologies that will be used for the automatic description of TROMPA music material. Although the main source of information are the **audio** itself, we are not limited to this, and we consider also **symbolic** and **video** descriptors which will be used to facilitate the use cases. Apart the technologies, we describe the potential use of these descriptors to the use cases. These descriptors will be deposited in various repositories related to the task needed and type of data and can be considered as contributions of TROMPA to public domain musical archives (e.g. automatic annotations, low-level descriptors). Moreover these data will also be linked to the Contributor Environment. Details on this process will provided in the next deliverables and **D5.3 - TROMPA Processing Library**[9] (M12) and **D2.3 - Technical Requirements and Integration**[10] (M18). The final version of this deliverable will be submitted in Month 20.

[7] https://trompamusic.eu/deliverables/TR-D5.3-TROMPA_Processing_Library_v1.pdf

[8] https://essentia.upf.edu/documentation/streaming_extractor_music.html

[9] https://trompamusic.eu/deliverables/TR-D5.3-TROMPA_Processing_Library_v1.pdf

[10] https://trompamusic.eu/deliverables/TR-D2.3-Technical_Requirements_And_Integration.pdf

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.

[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.

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