Deliverable D4.1

Report on the analysis and compilation of state-of-the-art methods for the automatic annotation of music pieces and music samples

Grant agreement nr	688382
Project full title	Audio Commons: An Ecosystem for Creative Reuse of Audio Content
Project acronym	AudioCommons
Project duration	36 Months (February 2016 - January 2019)
Work package	WP4
Due date	29 July 2016 (M6)
Submission date	29 July 2016 (M6)
Report availability	Public (X), Confidential ( )
Task leader	MTG-UPF
Authors	Dmitry Bogdanov, Xavier Favory, Frederic Font, Eduardo Fonseca
Document status	Final

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Table of contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Executive Summary

This deliverable provides an overview of existing state-of-the-art algorithms/descriptors for the automatic annotation of music samples and music pieces. The algorithms and music topics compiled in this deliverable have been chosen after examination of the output of the user survey included in deliverable D2.1 and with the aim of providing a general overview of common Music Information Retrieval (MIR) topics.

In this document we divide existing algorithms and descriptors in low-level and high-level categories. On the one hand, low-level descriptors include relatively simple representations of audio content (such as spectral representations) plus more complex musical representations such as those bearing information about tonality and rhythm. On the other hand, high-level (or semantic) features include musical concepts of a higher level of abstraction such as music genre or mood and which are typically computed using classifier models trained on annotated datasets or other machine learning techniques.

In this document, the emphasis is put on listing existing descriptors and relating them to available implementations rather than focusing on providing in depth algorithm descriptions and evaluations of their accuracy. For this reason, we include tables summarising existing algorithms and containing pointers to reference implementations as well as tables summarising datasets used for training and evaluation of particular tasks.

This document should be useful as a reference starting point for further research on feature extraction and high-level classification to be carried out in WP4 and, potentially, WP5.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Background

This deliverable belongs to the "Compilation of state-of-the-art methods for the automatic annotation of musical content" task (T4.1) of the "Semantic annotation of musical sound properties" work package (WP4). This document is partially based on the output of the deliverable "Requirements Report and Use Cases" (D2.1), in which a survey was carried out to learn what users expect of systems developed within AudioCommons. This document should serve as the basis for further research on automatic music annotation and for the final definition of the descriptors to be included in the music annotation tools that will be released within the AudioCommons project (deliverables D4.2, 4.3, 4.7, 4.8, D4.12 and D4.13).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

1 Introduction

1.1 Main objectives and goals

The goal of this deliverable is to provide an overview of existing state-of-the-art algorithms/descriptors for the automatic annotation of music samples and music pieces. The algorithms and music topics compiled in this deliverable have been chosen after examination of the output of the user survey included in deliverable Requirements Report and Use Cases (D2.1) and to provide a general overview of common Music Information Retrieval (MIR) topics.

1.2 Methodology

It is common practice in the MIR research field to make a distinction between low-level, mid-level and high-level audio features (or descriptors). In the "Roadmap for Music Information Research" published as a result of the MIReS EU project[1] (FP7, ICT-2011.1.5 Networked Media and Search Systems, grant agreement No 287711), the following definitions are proposed [Mires13]:

- ***Low-level features*** *relate directly to signal properties and are computed according to simple formulae. Examples are the zero-crossing rate, spectral centroid and global energy of the signal. Time-domain features such as the amplitude envelope and attack time are computed without any frequency transform being applied to the signal, whereas spectral features such as centroid, spread, flatness, skewness, kurtosis and slope require a time-frequency representation [...].*

- ***Mid-level features*** *(e.g. pitches and onset times of notes) are characterised by more complex computations, where the algorithms employed are not always successful at producing the intended results. Typically a modelling step will be performed (e.g. sinusoidal modelling), and the choice of parameters for the model will influence results. For example, in Spectral Modelling Synthesis, the signal is explained in terms of sinusoidal partial tracks created by tracking spectral peaks across analysis frames, plus a residual signal which contains the non- sinusoidal content [...].*

- ***High-level features*** *(e.g. genre, tonality, rhythm, harmony and mood) correspond to the terms and concepts used by musicians or listeners to describe aspects of music. To generate such features, the models employed tend to be more complex, and might include a classifier trained on a relevant data set, or a probabilistic model such as a hidden Markov model (HMM) or dynamic Bayesian network (DBN) [...].*

However, this distinction is not always clear as sometimes there are algorithms or music properties whose level of abstractness is ambiguous. For the sake of simplification, in this document **we will only make a distinction between low-level and high-level features**. On the one hand, low-level features include relatively simple representations of audio content (such as those included in the low-level and mid-level categories of the MIReS roadmap) plus more complex musical representations such as tonality and rhythm. On the other hand high-level (or semantic) features include musical concepts of a higher level of abstraction such music genre or mood and which are typically computed using classifier models trained on annotated datasets or other machine learning techniques.

This document is divided in two main sections which give an overview of the state-of-the-art of low-level and high-level features (sections 2 and 3 respectively). The **emphasis is put on listing existing**

[1] http://www.mires.cc

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

algorithms and relating them to available implementations rather than focusing on providing in depth algorithm descriptions and evaluations of their effectivity. For this reason sections 2 and 3 include tables summarising existing algorithms and containing pointers to reference implementations as well as tables summarising datasets used for training and evaluation of particular tasks.

1.3 Terminology

AudioCommons: reference to the EC H2020 funded project AudioCommons, with grant agreement nr 688382.

Audio Commons Initiative: reference to the AudioCommons project core ideas beyond the lifetime and specific scope of the funded project. The term "Audio Commons Initiative" is used to imply i) our will to continue supporting the Audio Commons Ecosystem and its ideas after the lifetime of the funded project, and ii) our will to engage new stakeholders which are not officially part of the project consortium.

Audio Commons: generic reference to the Audio Commons core ideas, without distinguishing between the concept of the initiative and the actual funded project.

Audio Commons Ecosystem (ACE): set of interconnected tools, technologies, content, users and other actors involved in publishing and consuming Audio Commons content.

Audio Commons content (ACC): audio content released under Creative Commons licenses and enhanced with meaningful contextual information (e.g., annotations, license information) that enables its publication in the ACE.

Content creator: individual users, industries or other actors that create audio content and publish in the ACE through content providers.

Content provider: services that expose content created by content creators to the ACE.

Content user: individual users, industries or other actors that use the content exposed by content providers and created by content creators in their creative workflows.

Tool developer: individual users, industries or other actors that develop tools for consuming (and also potentially publishing) Audio Commons content.

Embeddable tools: tools for consuming Audio Commons content that can be embedded in existing production workflows of creative industries.

Music samples: audio signals representing simple music events such as single notes, instrument hits, melodies and loops. Music samples can be used in music production contexts to, for example, build a sampler application using recordings of isolated instrument notes, or to create a music composition by overlaying different loops.

Music pieces: audio signals representing full compositions in the traditional sense of "songs". Music pieces can be used in a creative context as, for example, the soundtrack of a video or as remixable material for a music producer.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

2 Low-level descriptors

In this section we give an overview of existing low-level descriptors that can be automatically extracted from audio to annotate music pieces and music samples. The section is separated in a number of subsections that roughly relate to different music properties that can be annotated. Each subsection includes tables with links to existing implementations of some of the corresponding algorithms.

2.1 Dynamics

2.1.1 Loudness

According to Wikipedia, *Loudness is the characteristic of a sound that is primarily a psychological correlate of physical strength (amplitude). More formally, it is defined as "that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud"*[2]. In this document we consider loudness as a global single descriptor that would allow to order audio signals from quiet to loud. Different ways have been proposed for computing the loudness of an audio signal. Most basic ones are based on simple calculations over signals' energy (like using Steven's power law to measure the perceived intensity of a stimulus[3]). More complex ones take into consideration psychoacoustics and different time scales such as the recent EBU R128 recommendation for loudness measurement [EBU16]. A number of implementations of loudness algorithms can be found in the table below.

Table 2.1: Implementations of Loudness descriptors

Name/paper	Implementation	Link to code/repository
Loudness	Essentia (Loudness algorithm)	https://github.com/MTG/essentia/blob/master/src/algorithms/temporal/loudness.cpp
Vickers Loudness [Vickers01]	Essentia (LoudnessVickers algorithm)	https://github.com/MTG/essentia/blob/master/src/algorithms/temporal/loudnessvickers.cpp
EBU R128 [EBU16]	Essentia (LoudnessEBUR128 algorithm)	https://github.com/MTG/essentia/blob/master/src/algorithms/temporal/loudnessebur128.cpp
RMS (energy)	Essentia (RMS algorithm)	https://github.com/MTG/essentia/blob/master/src/algorithms/stats/rms.cpp
ReplayGain	Essentia (ReplayGain algorithm)	https://github.com/MTG/essentia/blob/master/src/algorithms/standard/replaygain.cpp https://en.wikipedia.org/wiki/ReplayGain
Loudness	LibXtract, also ported to a Vamp plugin	LibXtract: https://github.com/jamiebullock/LibXtract

[2] https://en.wikipedia.org/wiki/Loudness
[3] https://en.wikipedia.org/wiki/Stevens%27_power_law

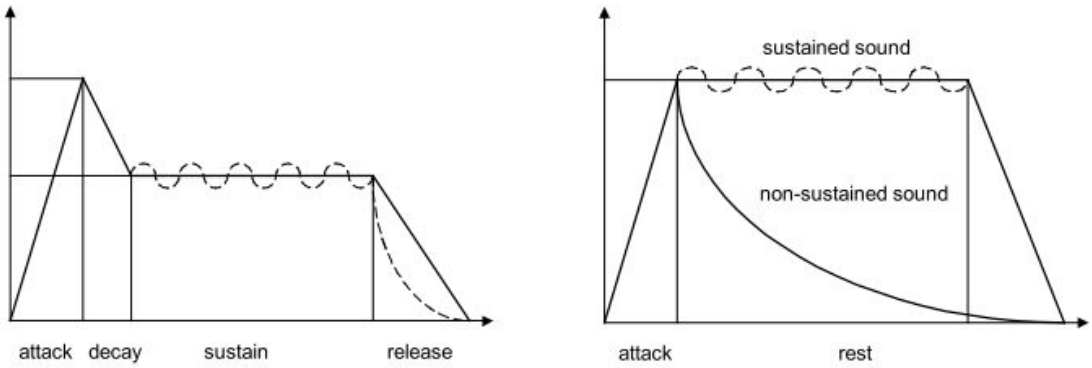
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

		Vamp plugin: https://code.soundsoftware.ac.uk/projects/vamp-libxtract-plugins

2.1.2 Envelope

By "envelope" we refer to the energy envelope of the signal, which approximates the temporal evolution of the loudness of an audio signal. As proposed in [Peeters04], the envelope of an audio signal can be characterised with the following descriptors:

- **Attack/decay/sustain/release (ADSR) related descriptors**: this is the way in which most synthesizers characterise the temporal envelope of generated sounds, dividing it in an attack, decay, sustain and release parts. The figure below (taken from [Peeters04]) shows a graphical representation of these parameters related to the different parts of the envelope (left). However, the ADSR model is typically simplified to an **Attack/Rest (AR)** model in which the attack is maintained as in the ADSR model but the 'rest' part can represent either a release for non-sustained sounds or a sustain for sustained sounds (see figure below, right). Note that, ADSR or AR are in fact **only useful to approximate the envelope of relatively simple audio events** like single hits, instrument notes or simple fx's, but are not very informative for music pieces, soundscapes or other complex audio signals.



To compute AR of an audio signal a thresholding approach can be used to estimate the start and end of the attack phase (i.e., points in time where energy goes above a minimum and maximum percentage of the maximum amplitude of the sound). Then, [Peeters04] propose **Temporal increase** and **Temporal decrease** descriptors as linear approximations of attack and rest parts of the envelope (i.e., before and after the end of the attack). **Log-Attack time** is another common descriptor that characterises the attack part and has been proven to be perceptually relevant. It is computed as the logarithm of the time elapsed between the start and the end of the attack [Peeters04]. For sustained sounds, the rest part can also be characterised by estimating its **Energy modulation** (tremolo).

- **Temporal centroid**: the temporal centroid is the *time averaged over the energy envelope*, and it *allows distinguishing percussive from sustained sounds* by approximating the point in time where energy is most concentrated [Peeters04]. Similarly to ADSR related descriptors, temporal centroid is not that meaningful to characterise sounds which very complex envelopes containing several events separated by silence parts.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

- **Effective duration**: the effective duration measures the time where an audio signal's energy is above a certain threshold [Peeters04]. This is the time where the signal is perceptually relevant. Similarly to previous descriptors, effective duration is only meaningful for characterising the envelope of single sound events.

The following table contains links to implementations of the algorithms above:

Table 2.2: Implementations of Envelope descriptors		
Name/paper	**Implementation**	**Link to code/docs**
Energy envelope, Temporal increase, Temporal decrease, Log-attack time, Energy modulation, Temporal centroid, Effective duration	Timbre toolbox (@cSound class)	http://www.cirmmt.org/research/tools
Energy envelope, Log-attack time, Effective duration	Essentia (Envelope, LogAttackTime and EffectiveDuration algorithms)	https://github.com/MTG/essentia/blob/master/src/algorithms/standard/envelope.cpp http://github.com/MTG/essentia/blob/master/src/algorithms/sfx/logattacktime.cpp http://github.com/MTG/essentia/blob/master/src/algorithms/temporal/effectiveduration.cpp
Attack start time, Attack end time, Log-attack time	MIR.EDU, also ported to a Vamp plugin	https://github.com/MTG/miredu/blob/master/src/AttackStartEndTimes.cpp https://github.com/MTG/miredu/blob/master/src/LogAttackTime.cpp

2.2 Pitch and Melody

Several algorithms in audio and music signal processing exist related to the estimation and characterization of a signal's tonal and melodic behaviour. Traditionally, an estimate of the pitch of a monophonic[4] music signal has been a very demanded feature. Also, the extraction of pitches of the multiple sources simultaneously present in polyphonic recordings has been a relevant topic. More recently, algorithms tailored to the extraction of the predominant source's pitch or melody in a mixture of other sources have been object of research.

Although these tasks are strongly linked, they present a different problem formulation and hence they can be treated separately. As will be explained later, melody extraction differs from monophonic and

[4] Monophonic signals can be defined as those when at most one note is sounding at a time. On the contrary, complex music signals where several sounds are played simultaneously are denoted as polyphonic signals [Klapuri06a].

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

polyphonic pitch estimation in two aspects [Salamon14]. On the one hand, the signal under consideration is polyphonic unlike in the monophonic pitch estimation problem. On the other hand, melody extraction entails the identification of a single source carrying the melody, whereas in the polyphonic pitch estimation task, multiple pitch values for several sources are outputted. We start by giving an overview of pitch estimation techniques for monophonic signals, continue with multiple pitch estimation in polyphonic music signals, and finish outlining melody extraction techniques for polyphonic music signals.

2.2.1 Pitch Estimation

Pitch may be defined as "that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale" [ANSI60]. Depending on the literature, the terms fundamental frequency, i.e., F0, and pitch are often interchanged. In most cases these two attributes coincide as pointed out by [Camacho08], although this statement does not always hold and counter examples can also be found [Patel01]. To be precise, the term fundamental frequency should be used to designate the physical property closely related to the subjective or perceptual attribute of pitch [Moore03]. Nonetheless, for the sake of simplicity and readability, we will use them indistinctly throughout this document.

Pitch estimation is a topic with long research tradition. A huge variety of algorithms have been proposed in the last few decades. For this document we have focused on approaches proposed mainly within the last decade and some other older methods that are relevant. Extensive research has been carried out in pitch estimation specifically for speech signals. In fact, some of the music oriented monophonic pitch estimators were adapted from preceding speech related algorithms. In this document, the object of study is pitch estimation methods that are somehow related to music content, either designed specifically for musical signals or at least evaluated with music data.

Monophonic pitch estimation techniques analyze the audio signal and attempt to detect the pitch from it. This techniques can be broadly categorized as follows [Gerhard03, Hajimolahoseini15]:

- **Time domain techniques**: methods that make use of the time-domain characteristics of the signal, e.g., autocorrelation function
- **Frequency domain techniques**: methods that rely on frequency-domain representations of the signal, e.g., spectrum or cepstrum, and
- **Parametric methods**, which use a signal model with parameters for frequency estimation.

Time-domain techniques estimate the fundamental frequency by analysing the signal waveform directly. Perhaps the most popular algorithm within this family is the YIN algorithm [DeCheveigne02]. This well known algorithm is based on the autocorrelation function with a series of improvements to reduce error rates. Recently, a modification of YIN (pYIN) has been proposed [Mauch14], following a probabilistic approach that enables to output multiple pitch candidates with associated probabilities for an improved pitch track. PRAAT algorithm, widely used for speech and phonetic analysis, is sometimes used as benchmark. PRAAT's default approach is based on an accurate autocorrelation function [Boersma93].

Among the frequency domain based algorithms is the YINFFT [Brossier07], an optimized version of YIN algorithm for computation in the frequency domain. Also, the SWIPE algorithm [Camacho08] estimates the pitch as the fundamental frequency of the sawtooth waveform whose spectrum best matches the spectrum of the input signal. One approach that can be conceptually regarded as a combination of the two approaches mentioned so far is the HPM algorithm [Luo14]. This method relies on autocorrelation both in the time domain and in the frequency domain.

Within the parametric family, a very recent work [Nielsen16] presents a maximum likelihood (ML) based estimator. The proposed method significantly reduces the computational cost of standard ML algorithms, while reporting higher accuracy than some autocorrelation based methods. Also, in

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

[Hajimolahoseini15] a so-called time-domain state-space model for the input signal is derived and supplied to several type of filters.

Other recent pitch estimation methods worth mentioning are briefly described next. In [Yang14], the proposed BaNa algorithm is claimed to provide high robustness on pitch detection performance with focus on speech and music in noisy environments. Despite the algorithm's public availability, the used analysis dataset is somewhat limited and more extensive tests would be desired to test its validity. An auditory inspired pitch extraction method is proposed in [Kumaresan14], emulating the different processing mechanisms of the auditory system involved in the perception of pitch. In [Babacan13], the adaptation of state-of-the-art speech oriented pitch extraction algorithms for the analysis of singing voice is discussed and a comparative evaluation of the most representative techniques is provided (PRAAT and YIN among others). Finally, a method to improve the accuracy of previously applied F0 estimators is proposed in [Stoter15]. It uses time warping iteratively based on an initial F0 estimate. The refinement can be applied to any F0 estimator and is not limited to specific input signal.

A list of publicly available implementations of algorithms[5] for pitch estimation in monophonic music signals is included in Table 2.3:

Table 2.3: Implementations of Pitch Detection algorithms

Name/paper	Implementation	Link to code/repository
YIN/ [DeCheveigne02]	Aubio (also as VAMP plugin) /Essentia	Aubio: http://aubio.org/manpages/latest/aubiopitch.1.html Aubio (vamp plugin): http://aubio.org/vamp-aubio-plugins/ Essentia: https://github.com/MTG/essentia/blob/master/src/algorithms/tonal/pitchyin.cpp
YINFFT/ [Brossier07]	Aubio/Essentia	Aubio: http://aubio.org/manpages/latest/aubiopitch.1.html Essentia: https://github.com/MTG/essentia/blob/master/src/algorithms/tonal/pitchyinfft.cpp
PYIN/ [Mauch14]	Vamp plugin	http://code.soundsoftware.ac.uk/projects/pyin
[Nielsen16]	MATLAB code	http://kom.aau.dk/~jkn/publications/publications.php
SWIPE/ [Camacho08]	MATLAB code	http://www.cise.ufl.edu/~acamacho/publications/swipep.m

[5]some algorithms are claimed to yield a high performance but have no implementations available. This is the case for the HPM algorithm [Luo14], reported to attain higher accuracy than SWIPE, YIN and AC-P algorithms for certain performance measures.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

BaNa/ [Yang14]	MATLAB code	http://www.ece.rochester.edu/projects/wcng/code.html
AC-P/ [Boersma93]	PRAAT	http://www.fon.hum.uva.nl/praat/

With respect to the music data used to evaluate the aforementioned pitch estimation algorithms, most works utilize a few instrument samples or synthetic signals used within the scope of the publication. However, some articles rely on part of the content of more extensive crafted databases. Table 2.4 lists datasets available to researchers that have been partly used for the evaluation of pitch estimation algorithms.

Table 2.4: Datasets for evaluation of monophonic pitch estimation systems

Dataset	Comments	Link to data
Musical Instrument Samples Database, University of Iowa Electronic Music Studios	Recordings of strings, woodwinds, brass, percussion and a Steinway piano, mostly carried out in anechoic chamber	http://theremin.music.uiowa.edu/MIS.html#
RWC Musical Instrument Sound Database [Goto03]	Isolated notes found among 150 instrument performances (3 variations each for 50 types of musical instruments)	https://staff.aist.go.jp/m.goto/RWC-MDB
MIREX[6] 2004	20 monophonic sound files from the MIREX 2004 melody extraction contest	http://ismir2004.ismir.net/melody_contest/results.html
MIDI Aligned Piano Sounds (MAPS) [Emiya10]	Among other types of content (chords and pieces), it contains isolated notes and monophonic sounds from a Yamaha Disklavier piano	http://www.tsi.telecom-paristech.fr/aao/en/2010/07/08/

2.2.2 Multi-Pitch Estimation

Multiple pitch estimation (MPE) consists of estimating the fundamental frequencies present in polyphonic sounds, typically over short time frames of an audio signal. In particular, in the Music Information Retrieval Evaluation eXchange (MIREX), the task related to Multiple Fundamental Frequency Estimation & Tracking is aimed at identifying "the active F0s in each time frame and to track notes and timbres continuously in a complex music signal". MPE is a challenging topic in polyphonic music transcription since it involves dealing with diverse spectral patterns in a wide frequency range that may

[6] Music Information Retrieval Evaluation eXchange (MIREX), http://www.music-ir.org/mirex/wiki/MIREX_HOME

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

include overlapping tones from multiple sources (harmonic or non-harmonic) and which quantity may be unknown.

According to [Benetos13a], MPE systems can be classified into three groups based on the core techniques employed:

- **feature-based approaches** use methods derived from signal processing where a specific model is not employed. Instead, the idea is to use measures of pitch salience from the input time-frequency representation and criteria for selecting pitch candidates.
- **statistical model-based approaches** formulate the multi-pitch estimation problem within a statistical framework, viewing it as as a maximum a posteriori (MAP) or as a maximum likelihood (ML) estimation problem.
- **spectrogram factorization-based approaches** utilize templates of spectral patterns of different pitch combinations and/or instruments, followed by input spectrogram decomposition based on template activation. This approach is followed by most recent works.

Within the first category, in [Klapuri06b] a set of possible fundamental frequency candidates is extracted from the spectrum based on the principle of harmonic summation. This is done iteratively, removing their harmonics from the mixture spectrum every time. Then, the number of harmonic sources is estimated and the final set of fundamental frequencies is determined in an optimization stage. In recent work [Su15a], a feature-based approach that leverages both frequency and periodicity aspects of music signals is proposed, referred to as combined frequency and periodicity (CFP). The idea is to exploit the fact that a pitch forms a harmonic series in the frequency domain and a subharmonic series in the lag (quefrency) domain. Pitches are detected according to the agreement between the two representations, thus aggregating the complementary advantages of the two domains.

An example of the second category is [Duan10], where multiple pitches are estimated with a maximum-likelihood approach that models both spectral peaks and non-peak regions of the power spectrum. It assumes spectral peaks at harmonic positions and lower energy elsewhere and proposes an iterative search strategy to estimate F0s one by one. A neighbourhood refinement method is used to eliminate inconsistent estimations and to refine the polyphony estimation.

Finally, some works that fall under the third category are described next. In [Vincent10], adaptive spectral decomposition is proposed using Non-negative Matrix Factorization (NMF) constrained by considering harmonicity and spectral smoothness. More specifically, basis spectra are modeled as weighted sum of narrowband spectra representing a few adjacent harmonic partials, adapting spectral envelopes to different instruments.

In [Benetos11], the Shift-Invariant Probabilistic Latent Component Analysis (SIPLCA) method is used, allowing for the detection of notes presenting non ideal tunings and frequency modulation patterns, and supporting multiple instrument models and pitch templates. Then, Hidden Markov Models (HMMs) are used in the post processing step to favour temporal continuity for note tracking. An improved version is presented in [Benetos13b], sharing the same time-frequency representation and note tracking steps, but additionally using spectral templates that correspond to sound states of the produced notes. This information is incorporated into the model by controlling the states' order through hidden Markov model-based temporal constraints.

Regardless the mentioned categories, in [Duan14] a method is proposed for the task of multi-pitch streaming, i.e., stream pitch estimates into trajectories over entire music performances for each of the concurrent sources. The proposed method takes the audio signal together with frame-level pitch estimates from any MPE algorithm as inputs. The output is a pitch trajectory for each source. The method follows a constrained clustering approach, aimed at minimizing timbre inconsistencies within every cluster, where each cluster corresponds to a source.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Table 2.5 lists some of the available open-source algorithms for MPE.

Table 2.5: Implementations of Multi-Pitch Estimation algorithms		
Name/paper	**Implementation**	**Link to code/repository**
[Klapuri06b]	Essentia	https://github.com/MTG/essentia/blob/master/src/algorithms/tonal/multipitchklapuri.cpp
[Benetos13b]	Vamp plugin	https://code.soundsoftware.ac.uk/projects/silvet
[Benetos13c]	MATLAB code	https://code.soundsoftware.ac.uk/projects/amt_mssiplca_fast
[Vincent10]	MATLAB code	http://www.irisa.fr/metiss/members/evincent/multipitch_estimation.m
[Duan10]	MATLAB code	http://www.ece.rochester.edu/~zduan/resource/Publications.html
[Duan14]	MATLAB code	http://www.ece.rochester.edu/~zduan/resource/Publications.html

Table 2.6 summarizes a number of datasets available to researchers for evaluating MPE systems. In certain works, only some excerpts from a dataset are used for evaluation, rather than the complete corpus.

Table 2.6: Datasets for evaluation of MPE systems		
Dataset	**Comments**	**Link to data**
MIREX 2007 MultiF0 Estimation Tracking Task	Development dataset used for the MultiF0 Estimation Tracking Task, including a woodwind quintet recording	http://www.music-ir.org/mirex/wiki/2014:Multiple_Fundamental_Frequency_Estimation_%26_Tracking
RWC Classical Music Database & Jazz Music Database [Goto02]	50 pieces for each database, of varied nature	https://staff.aist.go.jp/m.goto/RWC-MDB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Disklavier piano dataset [Poliner07a]	Includes 10 one-minute classical recordings from a Yamaha Disklavier grand piano	https://acousticbrainz.org/datasets/accuracy#genre_rosamerica
MIDI Aligned Piano Sounds (MAPS) [Emiya10]	Among other types of content (monophonic sounds and chords), it contains 30 music pieces played on Yamaha Disklavier piano	http://www.tsi.telecom-paristech.fr/aao/en/2010/07/08/
Bach10 [Duan10]	Includes recordings of ten pieces of J.S. Bach chorales performed by violin, clarinet, saxophone and bassoon	http://music.cs.northwestern.edu/data/Bach10.html
The TRIOS Score-aligned Multitrack Recordings Dataset [Fritsch12]	Contains five multitrack recordings of short musical extracts from trio pieces, featuring piano and 2 other pitched instruments	http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/27
Su dataset (MIREX 15) [Su15b]	Contains a wide range of real-world music: piano solo, string quartet, piano quintet and violin sonata	https://sites.google.com/site/lisupage/research/new-methodology-of-building-polyphonic-datasets-for-amt

2.2.3 Melody Extraction

Melody is regarded as one of the most relevant aspects of music, enabling a variety of potential applications: automatic transcription, melodic retrieval (query-by-humming) or as guidance for source separation techniques. Even though there does not seem to be a standard way to define *melody*, a commonly accepted definition in the MIR community is *the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music, and that a listener would recognize as being the 'essence' of that music when heard in comparison* [Poliner07b].

Typically, research has focused on single source predominant fundamental frequency estimation, where the term predominant refers to the source with the highest energy, be it from the lead voice or instrument in the recording. Therefore, we can say that melody extraction methods aim to obtain a sequence of frequency values corresponding to the pitch of the dominant melody from a musical recording [Salamon14]. More specifically, these systems are expected to accomplish two goals: estimate the correct pitch of the melody and the time intervals when the melody is actually present.

A list of some of the most relevant algorithms for melody extraction from polyphonic signals that have participated in MIREX up to 2012 can be found in [Salamon14]. Melody extraction algorithms can be considered as extensions of monophonic pitch trackers. In short, a pitch tracker takes an input signal, computes a function evaluated across a number of pitch candidates and finally some constraints are applied to produce the final pitch sequence. Thus, melody extraction algorithms can be classified according to their underlying approach [Salamon14]:

- **Salience-based approaches** consist of improving the robustness of the pitch candidate function against the rest of existing periodicities.
- **Source separation approaches** attempt to decompose the music signal into different sources in such a way that one of them is dominated by the melody signal. This newly obtained signal can input a simpler pitch tracker to obtain the melodic line.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

- **Other approaches** that do not fit into either category.

Most proposed melody extraction methods belong to the first group, i.e., they are based on the estimation of a **pitch salience function**. The main stages of the general architecture common to the majority of them are described next.

1. **Preprocessing**. Some works apply filters so as to enhance the frequency range where the melody is expected, e.g., [Salamon12], and/or a source separation method to enhance the melodic line and attenuate the accompaniment before the rest of the stages [Yeh12].
2. **Spectral Transform and Processing**. After frame-windowing the signal, a spectral representation for each frame is obtained using a number of different transforms, ranging from straightforward STFT [Salamon12] to other more sophisticated techniques, e.g., constant-Q transform [Cancela08]. Once in the spectral domain, the majority of the methods only keep the spectral peaks for further processing.
3. **Salience function.** The core of salience-based algorithms is the salience function. The aim is to create a time–frequency representation of pitch salience where prominence is assigned to each possible candidate pitch value on a frame by frame basis. For this purpose, most methods use some form of harmonic summation [Salamon12, Dressler12]. Nonetheless, other methods exist, e.g., [Durrieu11] which uses an approach based on NMF.
4. **Tracking.** Having obtained the salience function, the objective is to determine which peaks belong to the sought melody. Most works tend to track the melody directly from the salience peaks while in other studies a preliminary grouping stage is included, leading to pitch streams or contours [Cancela08, Salamon12]. In any case, several tracking techniques exist to select the final melodic sequence, e.g., [Durrieu10a, Bittner15].
5. **Voicing.** Last stage consists of detecting whether the melody is present or not on a frame by frame basis. Most algorithms rely on some static or dynamic thresholds [Durrieu10a, Fuentes12]. Others make use of pitch contour salience distributions [Salamon12] or the contour probabilities produced by discriminative models [Bittner15].

As previously indicated, **source separation** based approaches present an initial melody separation stage followed by pitch estimation and voicing detection. The idea is to isolate the melody source as much as possible making use of source separation techniques. An example of this approach is [Fuentes12], which uses a Probabilistic Latent Component Analysis to build a pitch salience function, followed by smoothing to estimate melody trajectory. In [Tachibana11], an extension of the Harmonic/Percussive Sound Separation algorithm is utilized.

The larger majority of methods fall within the two approaches mentioned so far. However, some **alternative strategies** can be found. That is the case of [Poliner05], relying on a machine learning scheme based on a support vector machine (SVM) classifier.

As a final comment to this section, it appears that melody extraction methods commonly yield better results for vocal music in comparison to instrumental pieces, since singing voice has been the main focus of study. According to [Bosch16b], this applies to the method proposed in [Salamon12], one of the best performing algorithms in terms of overall accuracy in past MIREX competitions.

Recently, research somewhat more oriented to the melody extraction of instrumental content has been carried out. In [Bosch16a], three melody extraction methods are proposed. In particular, different pitch tracking and voicing estimation techniques are combined with pitch salience computation based on source-filter modelling and harmonic summation. The proposed methods are based on combination and refinement of the processing blocks of some of the algorithms mentioned above [Durrieu10b, Bittner15, Salamon12] and a comparative evaluation with them is included. For the MedleyDB dataset, two of the proposed methods yield best overall accuracy while for the ORCHSET dataset, [Durrieu10b] attains best results. In light of the discussion, it appears that a different algorithm may be the most suitable option depending on the target musical content.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

In Table 2.7 we list some of the most important algorithms for automatically extracting the predominant melody from a polyphonic music recording that are publicly available.

Table 2.7: Implementations of Melody Extraction algorithms		
Name/paper	**Implementation**	**Link to code/repository**
[Bosch16a]	Python code	https://github.com/juanjobosch/SourceFilterContoursMelody
[Fuentes12]	Matlab code	http://www.benoit-fuentes.fr/articles/Fuentes2012_ICASSP/index.html
[Durrieu10b] [Durrieu11]	Python code	https://github.com/wslihgt/separateLeadStereo
[Bittner15]	Python code	https://github.com/rabitt/contour_classification
MEL-VAMP/ [Salamon12]	Vamp Plugin	http://mtg.upf.edu/technologies/melodia
MEL-ESS/ [Salamon12]	Essentia (algorithm PredominantPitchMelodia)	https://github.com/MTG/essentia/blob/master/src/algorithms/tonal/predominantpitchmelodia.cpp

A number of datasets specifically created to test the performance of melody extraction methods has been published. Table 2.8 lists some of the most relevant that are freely available.

Table 2.8: Datasets for evaluation of melody extraction systems		
Dataset	**Comments**	**Link to data**
ADC2004	20 excerpts in genres of pop, jazz and opera, including real recordings, synthesized singing and audio generated from MIDI files	http://labrosa.ee.columbia.edu/projects/melody/
MIREX05	13 excerpts in genres of rock, R&B, pop and jazz, including real recordings and audio generated from MIDI files	http://labrosa.ee.columbia.edu/projects/melody/
MedleyDB [Bittner14]	122 multitrack recordings (mix + processed stems + raw audio for music pieces and excerpts), where 108 of them have melody annotations	http://medleydb.weebly.com/
ORCHSET [Bosch16b]	64 audio excerpts from symphonies and symphonic poems, ballets suites and other musical forms interpreted by symphonic orchestras	http://mtg.upf.edu/download/datasets/orchset

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

MIR-1k dataset	1000 song clips from Chinese pop songs	https://sites.google.com/site/unvoicedsoundseparation/mir-1k
RWC Popular Music Database [Goto02]	100 popular music style songs with mainly Japanese lyrics and a few English lyrics	https://staff.aist.go.jp/m.goto/RWC-MDB/

2.3 Harmony

In this section, we discuss about harmony related annotations. Harmony refers to the use of different pitches simultaneously. First we tackle the problem of chord recognition systems and then discuss about key detection algorithms.

2.3.1 Chord recognition

Chord progression defines the harmonic structure in a music piece. Determining the chords within a musical composition is very important for more high-level analysis of the music signal. Chord grammar is the backbone of harmonic analysis and can give useful information for many applications. For instance, chord boundaries can be used for segmentation, music search and similarity identification. In a musical piece, pitches and chords are arranged according to a hierarchy of stabilities called tonality. The pitch class with the greatest stability is called the "tonic". The collection of chroma intervals with respect to the tonic form the “mode”. The combination of tonic and mode unambiguously describes a “key”. It is important to note that this musical system has been created for tonal Western music, and that in contemporary music for instance, the hierarchy between notes is often avoided and the tonal rules are not respected. Jazz music often uses tonal rules but tend to play more with seventh chords than triads.

An audio chord recognition system returns, from the audio signal, a sequence of labels indicating the chords in the piece with their activation times and durations. A typical chord recognition system consists of two main steps. First, the audio is cut into frames that are transformed in feature vectors. Then, each feature vector is mapped to a set of chord models. There are many ways to improve the chord recognition results. Enhancement techniques can be applied to the feature extraction or to the feature mapping.

Most of the chord recognition systems are based on chroma features. The chroma [Shepard64] is computed by summing the log-frequency spectrum across octaves. Then, the chroma frames are compared to the templates using a similarity or distance measure, and the best fit is determined. Recent works concentrate on improving the feature representation through a number of different techniques. Reassignment techniques in the spectrogram (called Time-frequency reassignment technique - TFR) remap spectral energy of each spectrogram cell into another cell in order to allow much higher time and frequency resolution [Khadkevich11].

For the classification part (e.g. the mapping step), there are many different ways for comparing chroma features and chord templates. Multiple distance metrics, such as cosine or euclidean distances are shown to yield good results [Fujishima99]. These methods provide a purely frame-wise recognition of chords, whereas other techniques take into account transitions between chords: Hidden Markov Model (HMM) has been widely used in many applications and is now the standard method in most automated procedures for chord recognition [Sheh03]. This method is very interesting because it provides a way to take into account the non arbitrary nature of chord progression. For example, some chord sequences are often used in a music piece or song (for example I-IV-V-I). The state transition probabilities can be specified manually (by music experts) based on harmony theory or automatically by estimating them from labeled data.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 688382

Other works use Conditional Random Field (CRF) to achieve the classification part [Wang13]. It differs from the HMMs in two things: First, is a discriminative model that theoretically is more suited for our problem, as it is not needed to model the probability of the data for a classification task. Additionally, in CRFs, each hidden state depends not just on the current observation but on the complete observation sequence. That gives the possibility to model the entire harmonic progression of music pieces where HMMs can only take into account local transitions between states.

In order to get around the problem of audio mis-tuning, some techniques can be used when extracting the pitch class profile. Khadkevich and Omologo [Khadkevich11] used two Fourier transforms of the signal, windowed by the same function but shifted by a number of samples. The phase change between the two spectra is used to calculate the frequency of harmonic components in a more accurate way.

Recent work utilize deep learning to learn alternative features for replacing chroma features [Zhou15]. In their work, authors investigate two types of architectures for the neural net, a common one in which the amount of neurons is the same in every layer, and a bottleneck-shaped architecture in which the middle layer has fewer neurons. Grézl et al. claim [Grézl07] that Bottleneck architecture is more suitable to learn high-level features than common one, and that it reduces overfitting. Moreover, following Zhou and Lerch results [Zhou15], it leads to better results for chord recognition than a common architecture.

The table 2.9 lists available implementations of chord recognition systems. We discuss these implementations below. Most of the systems use a vocabulary of chords containing only a subset of basic chords. For example, the chord estimation system implemented in Essentia takes into account only major and minor triads consisting in a vocabulary of 24 chords. Chordata from the CLAM project extend the vocabulary to more chords allowing the recognition of tetrads chords. Chordino allows user to define which chord can be recognized by letting him specifying chord profiles.

Essentia proposes a chord detection algorithm for offline or real time applications called ChordsDetection. Combined with the HPCP algorithm that computes the Harmonic Class Profile (HPCP), which is analogue of chroma features, from the frequency magnitudes of the audio signal, it forms a simple chord recognition system. However the modularity and simplicity of each block allows to build a more complicated system from it. Essentia also provides the ChordsDetectionBeats algorithm, which estimates chords on audio segments between each pair of consecutive beats. In addition, Essentia includes the ChordsDescriptors algorithm, which describes chord sequences by means of key, scale, histogram and rate of changes. This descriptor can be useful for computation of high-level descriptors such as genre. In addition, there is the TuningFrequency algorithm that estimates the tuning frequency of a song, given a set of spectral peaks. This tuning frequency is then passed on to the feature extraction algorithm (HPCP) as a parameter. It allows the algorithm to properly assign frequencies to chroma bins.

Chordino combined with NNLS Chroma forms a chord recognition system that seems to outperform Essentia implementation according to some informal experiments. Some pre-processing methods in the log-frequency spectrum give the system a way to be more robust and adjustable. Also two types of chromagram are outputted, one general-purpose that covers all pitches, and one bass-specific that is restricted to the lower frequencies. The classification step provided by Chordino algorithm is based on a simple chord transcription algorithm. Chord profiles given by the user are used to calculate frame-wise chord similarities. Two ways of smoothing the transcription are provided: a simple chord change method, and a standard HMM/Viterbi approach.

The LabROSA from Columbia University worked on Chord Recognition System. For MIREX 2010, Elis and Weller presented a system that uses instantaneous-frequency chroma features with a SVM classifier of HMM [Ellis10].

The CLAM project (C++ Library for Audio and Music) developed a software framework for research and application in the Audio and Music Domain. It offers many tools for the analysis, synthesis and processing of audio signals. Chordata is an application from CLAM that analyses the chords of any

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

music file. It provides some visualizations of the tonal features of the song. The processing algorithm inside the chord detection is based on Christopher Harte work [Harte05].

Alternatives of the use of chroma features for a chord recognition task has been studied. For example, autoencoders gave promising results [Steenbergen14].

Table 2.9: Implementations of Chord Recognition algorithms

Name/paper	Implementation	Link to code/repository
[Gomoez06]	Essentia algorithms: ChordsDetection, ChordsDetectionBeats, ChordsDescriptors, HPCP	https://github.com/MTG/essentia/blob/master/src/algorithms/tonal/chordsdetection.cpp
[Mauch10]	Vamp algorithms: Chordino NNLS Chroma	https://code.soundsoftware.ac.uk/projects/nnls-chroma/repository
HAMR 2014 Jeff Scott & Alex Cannon	Python	http://labrosa.ee.columbia.edu/hamr2014/proceedings/doku.php?id=simple_chord_recognition
[Steenbergen14] (master thesis)	Matlab	https://github.com/nikste/ChordRecognition
Gregory Burlet, 2012 (based on Neural Network and HMM)	Python	https://github.com/gburlet/chordRecog
CLAM Project	C++	http://clam-project.org/index.html
MusicBricks GenChords [Zen07]	C++	http://www.ifs.tuwien.ac.at/mir/chords/download/genchords.zip
Madmom [Korzeniowski16]	Python	https://github.com/CPJKU/madmom
MART-MIR [Khadkevich11]	Java	https://github.com/hutm/MART-MIR

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

In the following table 2.10, we list datasets that can be used to evaluate chord recognition systems.

Table 2.10: Datasets for evaluation of Chord Recognition systems

Dataset	Comments	Link to data
The Beatles	12 annotated albums of The Beatles. Often used by researchers	http://www.isophonics.net/content/reference-annotations-beatles
Zweieck	18 annotated tracks of Zweieck. Use with moderate confidence	http://www.isophonics.net/content/reference-annotations-zweieck
Queen	32 annotated tracks of Queen. Use with moderate confidence	http://www.isophonics.net/content/reference-annotations-queen
Carole King	7 annotated tracks of Carole King. Use with care	http://www.isophonics.net/content/reference-annotations-carole-king
Billboard	1000 professional chord transcriptions of popular music randomly selected from *Billboard's"*Hot 100" charts between 1958 and 1991 (no audio in the link)	http://ddmal.music.mcgill.ca/billboard
Robbie Williams	62 annotated tracks from the first five albums of Robbie Williams	http://ispg.deib.polimi.it/mir-software.html
Rock Corpus	500 Greatest songs from the Rolling Stone magazine (no audio in the link)	http://theory.esm.rochester.edu/rock_corpus/
UMA-Piano	275,000 annotated recordings of piano chords	http://extras.springer.com/2013/978-1-4614-7475-3
MIDI Aligned Piano Sounds (MAPS)	31 GB of CD-quality of piano recordings (contains isolated notes, chords and pieces of music)	http://www.tsi.telecom-paristech.fr/aao/en/2010/07/08/

2.3.2 Tonality

A key recognition system follows the same approach as chord recognition systems: First it extracts a set of low-level features from the audio signal (typically a global pitch class profile such as HPCP). Then these features are compared to a tonality model. As the tonality refers to a more global description of a music than chords, a global pitch class profile is computed by averaging instantaneous ones within a considered segment. So in general, the main difference with a chord recognition system is the use of a tonality model instead of chord templates.

There are different approaches for finding these tonality models. A simple way to do it is to start by the assumption that the most present pitch class in the piece is the tonic followed by the perfect fifth and the major or minor third.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

The Spiral Array graphical model[7] is more sophisticated and represent hierarchy for pitches, intervals, chords and keys. In this model, pitches are represented by point on a spiral. Adjacent pitches are related by intervals of perfect fifths from C (arbitrary chosen as the pitch reference). More information about such a model can be found in Chew's work [Chew01].

Other ways of constructing a tonality model would be to create key profiles by corpus analysis. This method is interesting because it allows the construction of genre specific models that can be more specialised and efficient for genre-restricted application. For example Faraldo et al. studied templated-based key estimation method for electronic dance music [Faraldo16].

Detecting key changes can be interesting for analyzing the structure of a music as pointed by Chain [Chai05]. In his work, he propose a HMM-based approach for segmenting musical signals based on key changes and identifying key of each segment. His result show the importance of taking into account key changes for the computation of key-adjusted self-similarity matrix that are commonly used for the automatic extraction of music structures.

The tables 2.11 and 2.12 below list available implementations of key detection algorithms, and datasets that can be used to evaluate them.

Table 2.11: Implementations of Key Detection algorithms

Name/paper	Implementation	Link to code/repository
[Gómez06]	Essentia algorithms: Key HPCP KeyExtractor	https://github.com/MTG/essentia/blob/master/src/algorithms/tonal/key.cpp
[Noland07]	Vamp algorithm: Key Detector	http://www.vamp-plugins.org/download.html?platform=linux64&search=key&go=Go
KeyFinder [Sha'ath11]	C++	https://github.com/ibsh/libKeyFinder

Table 2.12: Datasets for evaluation of Key Detection systems

Dataset	Comments	Link to data
The Beatles	12 annotated albums of The Beatles.	http://www.isophonics.net/content/reference-annotations-beatles

[7] https://en.wikipedia.org/wiki/Spiral_array_model

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Zweieck	18 annotated tracks of Zweieck. Use with moderate confidence	http://www.isophonics.net/content/reference-annotations-zweieck
Queen	32 annotated tracks of Queen. Use with moderate confidence	http://www.isophonics.net/content/reference-annotations-queen
Carole King	7 annotated tracks of Carole King. Use moderate confidence	http://www.isophonics.net/content/reference-annotations-carole-king
GiantSteps	604 annotated audio files from Beatport	https://github.com/GiantSteps/giantsteps-key-dataset
Robbie Williams	62 annotated tracks from the first five albums of Robbie Williams	http://ispg.deib.polimi.it/mir-software.html
KeyFinder v2	1000 annotated tracks	http://www.ibrahimshaath.co.uk/keyfinder/KeyFinderV2Dataset.ods
Mike Henderson	119 annotated tracks with result estimation of some commercial key detection software	http://dubspot.s3.amazonaws.com/public_files/Key%20Detection%20Lab%20Report%20by%20Endo.xlsx

2.4 Rhythm

In this section we give an overview of rhythm-related descriptors that are commonly extracted from music signals. We start with tempo and beat-tracking and then continue with meter and other rhythmic descriptors.

2.4.1 Tempo and beat-tracking

Tempo estimation consists in the automatic determination of the "rate of musical beats in time" [Gouyon06], that is to say, in the identification of the rate at which periodicities occur in the audio signal that convey a rhythmic sensation. Tempo is typically expressed in beats per minute (BPM). Beat-tracking consists in the "extraction of beat times from musical audio signals" [Zapata14]. Both features are intrinsically related. Tempo is nevertheless typically used as a global descriptor for an audio signal while beat-tracking outputs points in time where beats take place and is not summarised in a single descriptor. Assuming a constant tempo, beat positions could be estimated given a tempo value and a tempo value could be estimated from beat positions with high reliability.

A significant number of works within the MIR research field have been focused on the task of tempo estimation. In these works, many algorithms have been proposed which typically share a number of commons steps. In general, tempo estimation algorithms are based on detecting onsets in an audio signal, either as a continuous function [e.g., Davies07, Oliveira10, Percival14] or as discrete events in time [e.g., Dixon01, Zapata14]. Then, a dominant period is extracted from the onsets either by analysing inter-onset intervals, using autocorrelation [Grosche11] or resonating filters [Klapuri06c]. This results in a number of tempo candidates from which one is chosen after applying different heuristics to decide

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

which is the most probable tempo (e.g., to decide tempo octave). Some approaches perform more complex operations such as analysing periodicities in different frequency bands [Gainza11, Wu14], performing source separation [Gkiokas12, Elowsson13], or using neural networks to learn features to use instead of usual onset information [Böck15].

Beat tracking algorithms (as well as tempo estimation algorithms) are typically based on the extraction of low-level features from the audio signal that capture changes in its energy or spectrum (novelty function) [Müller15]. From these energy changes note onsets can be derived and then it can be measured how hypothetic beat sequences match with the strong onsets. Following this idea, an optimal beat sequence can be obtained that matches signal's onsets. Beat-tracking systems can also be used to estimate tempo [e.g., Dixon01, Davies07, Grosche11, Gkiokas12, Zapata14]. Estimated beat-positions are also very useful for carrying out other feature extraction tasks such as tonal analysis or structure analysis [Müller15].

In the past years two works have been published on comparative studies of tempo estimation algorithms [Gouyon06, Zapata11]. These works evaluate several state-of-the-art algorithms against common datasets (see below). A recent paper which has been published in the context of AudioCommons [Font16] evaluates tempo estimation algorithms in the context of music loops. Besides that, MIREX includes contests for both tempo estimation and beat-tracking. These provide unified framework for evaluating tempo estimation and beat-tracking algorithms. The table below provides a list of available datasets for the tasks of tempo estimation and beat-tracking.

Table 2.13: Datasets for tempo estimation and beat-tracking

Dataset	Comments	Link to data
MIREX 2006	Tempo estimation, beat-tracking. Dataset collected by Martin F. McKinney (Philips) and Dirk Moelants (IPEM, Ghent University). Composed of 160 30-second clips in WAV format with annotated tempos and beat positions.	http://www.music-ir.org/evaluation/MIREX/data/2006/tempo/ (user: tempo, password: t3mp0, only test data available)
ISMIR 2004	Tempo estimation. 3199 tempo annotated audio clips (Loops, Ballroom and song excerpts).	http://mtg.upf.edu/ismir2004/contest/tempoContest/ (only Ballroom and songs seems to be available)
GTZAN	Tempo estimation. Annotations for 1000 audio tracks each 30 seconds long. It contains 10 genres, each represented by 100 tracks.	http://marsyasweb.appspot.com/download/data_sets/ (annotations only) Rhythm annotations: http://anasynth.ircam.fr/home/media/GTZAN-rhythm
ACM MIRUM	Tempo estimation. Crowd-sourced tempo annotations.	http://www.marsyas.info/tempo/
GiantSteps	Tempo estimation. 664 audio files from Beatport.	https://github.com/GiantSteps/giantsteps-tempo-dataset
Freesound Loops	Tempo estimation. 4000 loops downloaded from Freesound and with tempo annotated	https://github.com/ffont/ismir2016

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

	by their uploaders.	
Mazurka	Beat tracking. 367 Chopin Mazurkas, used in MIREX beat-tracking evaluations.	(does not seem to be available any more online)
SMC MIREX 2012	Beat tracking. 217 excerpts around 40s each, particularly challenging for beat tracking.	http://smc.inesctec.pt/research/data-2/
Beatles	Beat tracking. Several annotations for Beatles songs (181 songs).	http://isophonics.net/content/reference-annotations-beatles
Robbie Williams	Beat-tracking. Annotations of beat positions for 5 Robbie Williams albums.	http://maxzanoni.altervista.org/chord-tracking/

In the context of the AudioCommons project we are particularly interested in experimenting with confidence measures that can estimate the reliability of audio properties annotated by the algorithms. Surprisingly enough, there has not been much research on confidence measures for tempo estimation and beat-tracking algorithms. Zapata et. al. [Zapata12] propose a confidence measure that can be used for tempo estimation and that is based on computing the mutual agreement between an ensemble of tempo estimation algorithms that take different sets of input features. Also, Grosche and Müller [Grosche11] describe a confidence measure for their tempo estimation algorithm based on the amplitude of a predominant local pulse curve (PLP curve). The PLP curve represents periodicities in the signal by fitting sinusoidal kernels to the novelty curves of each analysis window and performing an overlap-add operation. Tempo inconsistencies in the signal produce destructive interferences in the PLP, resulting in peaks of less amplitude. The confidence is then defined by setting an amplitude threshold for the PLP curve and selecting regions whose corresponding PLP peaks are above that threshold. By analysing tempo estimation accuracy and disregarding the regions of the analysis with bad confidence, the overall accuracy significantly increases. Alternatively, Percival and Tzanetakis [Percival14] suggest that beat strength [Tzanetakis02a] can be used to derive confidence for tempo candidates, but no further experiments are carried out to asses its impact on the accuracy of tempo estimation. Finally, a very recent work by Quinton et. al. [Quinton16] proposes the use of rhythmogram entropy as a measure of reliability for a number of rhythm features, and report a statistical correlation between measured entropy and the resulting accuracies for different tasks.

To finish this section, the table below lists existing open implementations for some of the tempo estimation and beat-tracking algorithms mentioned above.

Table 2.14: Implementation of tempo estimation and beat-tracking algorithms		
Tempo		
Name/paper	**Implementation**	**Link to code/docs**
[Böck15]	Madmom, "TempoDetector --method comb" utility	https://github.com/CPJKU/madmom
[Percival14]	Marsyas	https://github.com/marsyas/marsyas/tree/master/scripts/large-evaluators/tempo-reference-implementation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Beat-tracking (and tempo based on beat-tracking)		
Name/paper	**Implementation**	**Link to code/docs**
[Degara12]	Essentia (RhythmExtractor2013 algorithm)	https://github.com/MTG/essentia/blob/master/src/algorithms/rhythm/rhythmextractor2013.cpp
[Zapata12]	Essentia (RhythmExtractor2013 algorithm)	https://github.com/MTG/essentia/blob/master/src/algorithms/rhythm/rhythmextractor2013.cpp
Beatroot [Dixon01]	Vamp plugin (only beat tracking)	https://code.soundsoftware.ac.uk/projects/beatroot-vamp
[Gkiokas12]	Matlab code	Obtained by request to authors, hosted (with permission) at https://github.com/ffont/ismir2016/tree/master/algorithms/Gkiokas12
[Grosche11]	Tempogram toolbox/VAMP plugin implementation of tempograms	Tempogram toolbox: https://www.audiolabs-erlangen.de/resources/MIR/tempogramtoolbox/ VAMP plugin: https://code.soundsoftware.ac.uk/projects/vamp-tempogram/repository
[Oliveira10]	Marsyas	http://marsyas.info/doc/manual/marsyas-user/ibt.html#ibt
[Davies07]	Vamp plugin	http://vamp-plugins.org/plugin-doc/qm-vamp-plugins.html#qm-tempotracker

2.4.2 Meter and other rhythm descriptors

According to Wikipedia, meter describes music's rhythmic structure, "the patterns of accents heard in regularly recurring measures of stressed and unstressed beats [..]". Depending on the time scale at which meter is analysed, Klapuri et al. [Klapuri06c] propose to differentiate between *tatum* (corresponding to the lowest metrical level, i.e., the shortest notes [Seppanen01]), the *tactus* (which corresponds to the most dominant metrical level, usually maintaining an integer relation with the tatum and defining the tempo of a signal), and the *measure* (which tends to relate to harmonic and rhythmic changes). Meter is typically represented in music scores and production software using time signatures (e.g., 4/4, 6/8).

Much of the existing work on automatically deriving meter from music is based on the the analysis of MIDI files (i.e., symbolic music representations) rather than raw audio. Within AudioCommons we are interested in the determination of meter information directly from the audio signal. Some approaches for meter estimation consist in analysing the peaks of a periodicity function. Similarly to tempo estimation, a peak should be observed at the beat rate but other peaks should be observed which harmonic relations from which time signature could be estimated [Brown93]. Also, given all peaks of a periodicity function, a probabilistic approach can be followed to find the pair of peaks which is more likely to

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

represent beat and measure of a time signature [Dixon03]. Other approaches can take advantage of other features such as low-level spectral features or harmonic features to determine changes at the measure level for a given tempo [Goto99, Gouyon03]. Gainza [Gainza09] computes a beat-beat similarity matrix to pairwise compare all beats of a piece of music and find regular patterns that suggest a particular meter structure. More recent work proposes the detection down-beats (typically first beats of a measure) as an additional feature which can lead to the identification of a time signature [Krebs13], and beat-synchronous spectrum analysis to put emphasis on the temporal evolution of features [Varewyck13].

Besides tempo, beat positions and meter there are other rhythmic properties that could be automatically extracted from audio and that would be clearly relevant for its reuse in a creative scenario. And example of such rhythmic descriptors is the "loopability" of a signal (i.e., whether the signal can be seamlessly repeated to produce a continuous musically meaningful output). Roma [Roma15] proposes an approach based on analysis of the beat spectrum to automatically determine whether a given audio signal is loopable or not. The main idea is to check whether one of the prominent peaks of the beat spectrum is an integer multiple of the duration of the signal, thus indication that there is an harmonic relation between the duration and some rhythmic properties of the signal.

The following tables 2.13 and 2.14 include a relation of available algorithms and datasets for meter detection and other rhythm features. Unfortunately, not many algorithms and datasets are available when compared to other music properties.

Table 2.15: Implementation of meter and other rhythmic descriptors

Name/paper	Implementation	Link to code/docs
Downbeat tracking	Madmom, "DBNDownBeatTracker" utility	https://github.com/CPJKU/madmom
Rhythm Transform [Guaus05]	Essentia. Computes periodicity function based on several descriptors.	http://essentia.upf.edu/documentation/reference/std_RhythmTransform.html

Table 2.16: Datasets for meter and other rhythmic descriptors

Dataset	Comments	Link to data
Ballroom dataset beat and bar annotations	Down-beat/beat tracking. Adds complementary annotations to the Ballroom section of ISMIR 2004 dataset (see Table 2.13).	https://github.com/CPJKU/BallroomAnnotations
HJDB	Downbeat annotations. 235 excerpts of Hardcore, Jungle and Drum and Bass music between 30s and 2 minutes in length (all 4/4 time signature).	http://ddmal.music.mcgill.ca/breakscience/dbeat/
Carnatic Music Rhythm dataset	Annotated test corpus of 176 carnatic Indian music pieces (includes meter annotations).	http://compmusic.upf.edu/carnatic-rhythm-dataset

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

2.5 Other descriptors

So far we have given an overview of algorithms and datasets for the most common musical properties that can be described about sounds (envelope, pitch, harmony and rhythm). There are nevertheless many other aspects and descriptors that can be used to characterise music samples and music pieces. Another clear set of descriptors would be those related to timbre. The most basic amongst these include low-level descriptors that characterise the spectrum of audio signals. For example, Mel-frequency cepstral coefficients (MFCCs) are one of the most common and widely used low-level representations of the spectrum with applications including speech recognition and MIR classification tasks. These descriptors are not strictly musical properties in the sense that they do capture information with a perceivable meaning both for musical and non-musical audio content. In the context of the AudioCommons project, timbre descriptors will be dealt with in the different tasks and deliverables of WP5, therefore these are out of the scope of the present document.

It can be argued that further musical aspects that can be characterised from music samples and pieces bear a higher semantic meaning. That would include concepts like genre, mood or instrumentation of a particular recording. Some of these audio descriptors are dealt with in the following section (Sec. 3).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

3 High-level/Semantic descriptors

In this section we overview existing approaches to high-level/semantic annotation of music pieces and music samples mainly based on classification. The section is separated in a number of subsections covering different musical properties and includes tables with links to existing implementations of some of the corresponding algorithms and datasets for training and evaluating models.

As we have noted in Section 1.2 (Methodology), we distinguish high-level annotations from low-level features, the former being musical concepts of a higher level of abstraction such music genre or mood and which are typically computed using classification or other machine learning techniques.

Currently, when used in many practical applications such as music recommendation and music similarity, low-level and mid-level music features have shown a detrimental performance when compared with approaches working with metadata [Slaney11]. This might be explained by the fact that low- and mid-level music features extracted from music audio are close to the signal, but far from the way listeners do conceptualize or think about music. In contrast, it might be desirable to work on a higher level of automatic description and utilize semantic concepts used by humans. It is challenging to try to bridge the so-called semantic gap [Aucouturier09, Celma06, Fu11], which arises from the weak linking between human concepts related to musical aspects and the low-level features.

Currently, industrial systems employ manual annotation by semantic concepts done by music experts which significantly complicates the scalability of these systems but ensures the quality of annotations. As an example, Jamendo[8] provides expert annotations for a number of semantic concepts to enhance search functionality within its digital music library (genres and sub-genres, themes and moods, instruments, and languages). Similarly, Pandora music streaming service bases its recommendation engine on more than 450 semantic descriptors[9] annotated by musicologists.

Automatic classification and auto-tagging are therefore common problems considered by MIR researchers to address the challenge of bridging the semantic gap [Schedl14]. According to Wikipedia, *"Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known"*.[10] Typically many MIR classification tasks consider the task of assigning a single category to unknown music pieces. Nevertheless, annotation with multiple categories is also possible. The latter is typically known as auto-tagging. Commonly, the approaches for classification and auto-tagging are supervised and therefore they rely on annotated training datasets.

A number of relevant MIR tasks include:

- genre classification [Tzanetakis02, Scaringella06, Bogdanov16]
- music emotion / mood classification [Yang08, Laurier10]
- music culture classification [Gómez09]
- music instrument classification [Herrera06]
- music/speech classification and detection [Ghodasara15]
- semantic auto-tagging [Sordo12]

Classification can be performed on either segment or complete track level. The former includes instrument recognition, where instrumentation may vary from segment to segment, while the latter is commonly performed for the majority of classification tasks such as genre and mood where a single annotation per track is required [Fu11].

[8] https://licensing.jamendo.com/en/catalog
[9] https://en.wikipedia.org/wiki/Music_Genome_Project
[10] https://en.wikipedia.org/wiki/Statistical_classification

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Classification process is based on the results of low-level/mid-level feature extraction and typically requires preliminary steps of statistical summarization and pre-processing (normalization, feature selection). Fu et al. [Fu11] provides an overview of audio-based features describing timbre, temporal evolution, rhythm, pitch and harmony, commonly used for music classification. It is noted that the choice of audio features is much dependent on the concrete classification problem. Support vector machines (SVMs) are an effective and commonly used state-of-the-art tool for various classification tasks in MIR [Xu03, Mandel05, Laurier09, Bogdanov11]. Apart from SVMs, researchers have also employed other classifier models including:

- K-nearest neighbor (KNN) [Schnitzer11, Sordo12],
- Gaussian Mixture Models [Slaney02, Barrington07],
- Neural networks [Berenzweig03, Shawe-Taylor05, Hamel09]),
- Random forest [Kursa09]
- Naive bayesian [Zhang09]
- Ensemble algorithms (AdaBoost [Bergstra06, Yang08], bagging [Yah14])
- HMMs [Chai01]

It is worth to mention that deep convolutional neural networks have gained a large interest of the research community in recent years and have been applied for music classification task. Nevertheless they have not been demonstrated any improvement over the existing state of the art except for music instrument detection [Han16].

It should be noted that classifier models should be trained on many examples (music pieces or music samples) in order to ensure training sample representativity and a proper between/within class variation. Insufficient training dataset sizes or insufficient coverage can lead to overfitting which results in that the trained models are inadequate for new data. This is especially important in the case of high dimensionality of the input data, that is, when the number of low-level and mid-level features is comparable or higher to the number of training instances.

Importantly, the methodology of evaluation of classifier models has started evolving within MIR community within the last years. Many researchers limit their evaluations to cross-validation on small-sized datasets available to the community. However, this leaves the question of the practical value of these classifier models for annotation, if the goal is to apply a label to any unknown musical input. It has been shown that despite high cross-validation evaluation results trained models may work with irrelevant features and therefore learn something irrelevant to human understanding of music [Sturm14]. Some studies have noted that cross-validation may be prone to biases and that additional evaluations based on independent out-of-sample data are desirable [Bogdanov16]. This further motivates the creation of annotated datasets from various sources, including expert annotation, social tags (folksonomies) and other data mined from the Internet [Porter16], as well as the improvement of low- and mid-level audio descriptors used by machine learning approaches. Recent research effort is focused on creation of an open web platform for collaborative gathering and annotation of datasets of low- and mid-level audio features for various classification tasks[11] [Porter15].

We discuss specific approaches and datasets for semantic annotation below. In addition to purely machine learning approaches working with information resulting from audio analysis (*features*), many researchers derive audio feature sets designed specially for classification tasks. Such approaches therefore require access to audio collections in contrast to purely machine learning approaches which can be purely based on datasets of pre-computed common audio features.

Table 3.1 summarizes existing MIR software tools and some of the commonly used machine learning tools that can be used for semantic annotation tasks:

[11] https://acousticbrainz.org

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Table 3.1: MIR and machine learning tools for semantic annotation			
Tool	**Techniques**	**Interface**	**Link to code/documentation**
Essentia/Gaia [Bogdanov13]	Classification (SVM, nearest neighbour). Includes pre-processing, feature selection and dimensionality reduction. Includes pre-trained classifier models.	C++, Python, CLI	https://github.com/MTG/essentia/blob/master/src/algorithms/highlevel/gaiatransform.cpp https://github.com/MTG/gaia
jMIR [McKay09]	Classification (SVM, Naive Bayesian, decision trees, backprop neural network), ensemble algorithms (AdaBoost, Bagging). Includes pre-processing, feature selection and dimensionality reduction.	Java, CLI, GUI	http://jmir.sourceforge.net/manuals/ACE_manual/ACEManual.html
Marsyas [Tzanetakis00]	Classification (SVM, Naive Bayesian). Includes pre-processing and dimensionality reduction.	C++, CLI	http://marsyasweb.appspot.com/assets/docs/manual/marsyas-user/kea.html#kea
openSMILE [Eyben16]	Classification (SVM). Includes pre-processing, feature selection.	C++, CLI	http://audeering.com/research/opensmile/#opensmile
pyAudioAnalysis [Giannakopoulos15]	Classification (SVM, nearest neighbour), regression. Includes pre-processing.	Python	https://github.com/tyiannak/pyAudioAnalysis
AcousticBrainz [Porter15]	Classification (SVM). Includes pre-processing, feature selection.[12] Includes pre-trained classifier models.	Web	https://acousticbrainz.org/datasets/accuracy

[12] Web interface for specifying pre-processing and feature selection is currently under development.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Scikit-Learn	Classification (SVM, nearest neighbors, random forest, decision trees), regression, HMM. Includes pre-processing, feature selection and dimensionality reduction.	Python	http://scikit-learn.org
WEKA [Hall09]	Classification (many), clustering (many), regression (many). Includes pre-processing, feature selection.	Java, CLI, GUI	http://www.cs.waikato.ac.nz/ml/weka/
Shogun Machine Learning Toolbox [Sonnenburg10]	Classification, regression, HMM.	C++, CLI, Python, MATLAB, Java, R	http://www.shogun-toolbox.org

In the following sections we outline research and datasets about musically-relevant semantic concepts including genre, mood/emotion and instrumentation.

3.1 Genre

Genre classification is a common MIR task as it is believed to showcase the possibility of practical applications of MIR systems despite the somewhat subjective nature of the classification. It is therefore a recurring task at the annual Music Information Retrieval Evaluation eXchange (MIREX).[13]

Table 3.2 lists datasets available to researchers for training models for genre classification. Notably, some of the existing datasets (GTZAN and Latin Music Database) were largely criticized for a number of flaws, including insufficient coverage, conflicting annotations, and present of duplicate training instances [Sturm12, Sturm15].

Recent cross-cross evaluation revealed poor performance of existing genre classifier models trained on common small-sized MIR genre datasets, when applied for large-scale annotation on "real" music collections [Bogdanov16]. Poor performance can be explained by insufficient low- and mid-level information, not capturing the essence of various music genres as well as by ill-defined taxonomy of music genres and the representativeness of the associated datasets. Studies reveal disagreement of genre annotations between different sources due to a disagreement in the underlying taxonomies [Scaringella06], and further research should be focused on defining a common genre taxonomy with well-defined categories.

[13] http://www.music-ir.org/mirex/wiki/2015:Audio_Classification_%28Train/Test%29_Tasks

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Table 3.2: Datasets for genre classification		
Dataset	**Comments**	**Link to data**
Datasets including audio		
GTZAN [Tzanetakis02b]	Includes audio retrieved from personal collections. 1000 track excerpts (30s.) for 10 broad genres (100 per genre).	http://marsyasweb.appspot.com/download/data_sets
Music Audio Benchmark Data Set (MABDS) [Homburg05]	Includes audio retrieved from an online music community. 1886 full tracks for 9 broad genres (46–490 per genre).	http://www-ai.cs.uni-dortmund.de/audio.html
AcousticBrainz Rosamerica [Guaus09]	In-house audio collection owned by MTG created by a professional musicologist. 400 full tracks, 8 broad genres (50 tracks per genre).	https://acousticbrainz.org/datasets/accuracy#genre_rosamerica
Latin Music Database [Silla08]	Includes audio retrieved from personal collections and annotated by professional latin dance teachers. 3160 full audio tracks, 10 latin music genres.	Currently offline
MIREX Audio US Pop Music Genre Classification	In-house audio collection held by MIREX for blind evaluations. 7000 track excerpts (30s.) for 10 genres (700 per genre).	http://www.music-ir.org/mirex/wiki/2015:Audio_Classification_%28Train/Test%29_Tasks#Audio_US_Pop_Music_Genre_Classification
Datasets including only audio features		
MSD Allmusic Genre and Style Datasets (MAGD and MASD) [Schindler12]	Genre and music style annotations of the Million Song Dataset, derived from AllMusic expert annotations. 433,714 tracks annotated by 21 broad genres; 307,790 tracks annotated by 210 music styles from AllMusic genre taxonomy.	http://www.ifs.tuwien.ac.at/mir/msd/download.html
Tagtraum genre annotations [Schreiber15]	BeaTunes Genre Dataset (BGD): genre annotations for Million Song Dataset derived from beaTunes social tags (609,865 annotated tracks). Last.fm Genre Dataset (LFMGD): genre annotations for Million Song Dataset derived from Last.fm social tags (340,323 annotated tracks). Combined annotations using agreement between sources (133,676 - 280,831 annotated tracks).	http://www.tagtraum.com/msd_genre_datasets.html
AcousticBrainz /	MAGD annotations mapped to AcousticBrainz	http://labs.acousticbrainz.org

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

MSD Allmusic Genre Dataset (AB-MAGD) [Bogdanov16]	collection of audio features (142,969 tracks, 11 genres).	
AcousticBrainz / Tagtraum genre annotations [Bogdanov16]	Combined Last.fm/beaTunes genre tag annotations mapped to AcousticBrainz collection of audio features (148,960 tracks, 13 genres).	http://labs.acousticbrainz.org
AcousticBrainz genre annotations [Bogdanov16, Porter15]	MAGD annotations mapped to AcousticBrainz collection of audio features (142,969 tracks, 11 genres). Combined Last.fm/beaTunes genre tag annotations mapped to AcousticBrainz collection of audio features (148,960 tracks, 13 genres). 764,555 tracks annotated by AllMusic genre tree (21 top-level genre, 1186 genres). 720,597 tracks annotated by Discogs genre tree (15 top-level genres, 491 genres). 957,529 tracks annotated by Itunes genre tree (38 top-level genres, 253 genres). 841,571 tracks annotated by Last.fm social tags mapped to Discogs genre tree. 810,655 tracks annotated by Last.fm social tags mapped to AllMusic genre tree. 788,426 tracks annotated by Last.fm social tags mapped to Itunes genre tree. 778,964 tracks annotated by Last.fm social tags mapped to Beets genre tree (16 genres).	http://labs.acousticbrainz.org

3.2 Mood / Emotion

Classification of music by mood/emotion is another important MIR task present in MIREX. The purpose is to classify music pieces or samples into different emotional categories like happy, sad, angry, relaxed, etc. [Laurier09] Similarly to genre, music moods are difficult to infer as people perceive them differently [Song13] and they are culture dependent. In contrast studies in psychology have proposed two-dimensional representation of emotion using the arousal-valence plane [Russel80] and therefore some MIR research works are focused on a somewhat simpler task of inferring arousal/valence values. To this end, linear regression [Eerola09], neural networks [Coutinho14] and support vector regressions [Han09] have been used to map low- and mid-level features to arousal/valence values.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Table 3.3 summarizes datasets available to researchers for training mood/emotion models. In contrast to genre, large datasets based on existing collections of audio features (such as AcousticBrainz or Million Song Dataset) have not been yet established, although some research is ongoing [Corona15].

Table 3.3: Datasets for mood/emotion classification		
Dataset	**Comments**	**Link to data**
Datasets including audio		
MoodSwings Turk Dataset [Schmidt11]	Arousal/valence annotations gathered using Mechanical Turk for 240 analyzed music excerpts (15s.) (only audio features are provided).	http://music.ece.drexel.edu/research/emotion/moodswingsturk
emoMusic [Soleymani13]	Arousal/valence annotations for 744 music excerpts (45s.) from Free Music Archive.	http://cvml.unige.ch/databases/emoMusic
Soundtracks [Eerola11]	Collaborative annotations by valence, arousal, tension, and moods (anger, fear, sadness, happiness and tenderness) for 110 film soundtracks excerpts (10-30s.).	https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/emotion/soundtracks
Emotify [Aljanaki15]	400 song excerpts (1m.) collaboratively annotated by 9 induced emotion categories	http://www.projects.science.uu.nl/memotion/emotifydata
International Affective Digital Sounds (IADS)	Pleasure, arousal, dominance ratings for 111 sound snippets created for use in experimental investigations of emotion and attention	http://csea.phhp.ufl.edu/media/iadsmessage.html
MIREX Audio Mood Classification Dataset [Hu07, Downie08]	In-house audio collection held by MIREX for evaluations. 600 track excerpts collaboratively annotated by 5 mood clusters (120 per cluster).	http://www.music-ir.org/mirex/wiki/2015:Audio_Classification_%28Train/Test%29_Tasks
MIREX Audio K-Pop Music Classification	In-house audio collection held by MIREX for evaluations. 1894 k-pop song excerpts (30s.) in five mood categories, annotated by a number of American and Korean annotators.	http://www.music-ir.org/mirex/wiki/2015:Audio_K-POP_Mood_Classification
AcousticBrainz moods [Laurier09]	In-house audio collection owned by MTG. Contains annotations for: aggressive/non-aggressive (280 full tracks + excerpts), relaxed/non-relaxed (230 full tracks + excerpts), happy/non-happy (302 full tracks + excerpts), and sad/non-sad (230 full tracks + excerpts) categories done by a single person.	https://acousticbrainz.org/datasets/accuracy#mood_acoustic

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

3.3 Instrumentation

Instrument recognition is a challenging task that has not been yet addressed in MIREX but maintains a steady interest within MIR community. For monophonic sounds, such as pitched instrument samples including single notes or entire musical phrases, the recognition task is reduced to assigning a single instrument label. It is approached as a classification task, relying on a number of state-of-the-art audio features describing energy (RMS, loudness), spectral features (including spectral centroid, flatness, roll-off, skewness, kurtosis and spread, spectral flux, spectral irregularity, zero crossing rate and MFCCs), harmonic features (inharmonicity, odd-to-even ratio, tristimulus) and temporal evolution of a signal (amplitude envelope, attack time, temporal centroid, vibrato) [Herrera06]. Again, as in the case of other classification tasks, SVM is a common classifier of choice [Liu10].

A more complex task is instrument recognition in polytimbral music audio. It has not yet been addressed in MIREX but it maintains a steady interest within MIR community [Fuhrmann12]. The task is to output labels for all instruments present in a given polyphonic music piece, and also to locate their positions in time within the piece. Multiple instruments can occur in the piece in different times or simultaneously, and the analysis can be done involving classification on frame level, segments, or frame clusters [Fuhrmann11]. Fuhrmann [Fuhrmann11] proposes an approach to detect predominant instruments in polyphonic music by applying SVMs. Generally, the higher polyphony, the lower accuracy is usually obtained in automatic instrument recognition [Kubera14]. More recent approaches allow to improve the recognition accuracy by applying audio segregation techniques (panning-based segregation or more complex source separation techniques) [Bosch12]. Finally, deep convolutional neural networks have been also successfully used [Han16].

Classification requires development of instrument taxonomies and associated datasets (Table 3.4). Instrument families are easier to identify than particular instruments as for latter confusions are common even among human listeners. Similarly to genre, research on instrument classification includes examples of cross-collection evaluation in order to avoid possible biases associated with typical cross-fold validation [Livshin03].

Notably, existing approaches for instrument identification consider conventional instruments, while many other instruments are left out of scope. In particular, this applies for electronic music, some genres of which are heavily characterized by using particular type of synthesizers.

Table 3.4: Datasets for instrument classification

Dataset	Comments	Link to data
McGill University Master Samples collection (MUMS) [Eerola08]	6000 sound samples representing most classical and popular musical instruments and a wide variety of articulation styles. Contains labeling errors [Eerola08].	Available on DVD
RWC Musical Instrument Sound Database [Goto03]	150 instrument performances (3 variations each for 50 types of musical instruments)	https://staff.aist.go.jp/m.goto/RWC-MDB
Instrument Recognition from Audio Dataset	6705 polyphonic audio excerpts (3s.) for more than 2000 distinct music pieces annotated by 11 predominant instruments.	http://www.mtg.upf.edu/download/datasets/irmas

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

(IRMAS) [Bosch12]		
MedleyDB [Bittner14]	122 multitrack recordings (mix + processed stems + raw audio for music pieces and excerpts) annotated by instrument.	http://medleydb.weebly.com
Musical Instrument Samples Database, University of Iowa Electronic Music Studios	Recordings of strings, woodwinds, brass, percussion and a Steinway piano, mostly carried out in anechoic chamber	http://theremin.music.uiowa.edu/MIS.html#

3.3 Other semantic facets

These is a large interest to explore descriptors for other semantic facets, including specific timbral categories, rhythmic and tonal patterns, production techniques, speech/music classification, and structure of music pieces and music samples. While speech/music classification is relatively mature [Williams99, El-Maleh00, Ghodasara15] and has been included in the recent editions of MIREX, [14] other topics still lack further research. A number of semantic facets are covered within AcousticBrainz project providing datasets and classifier models for a few generic properties of the music pieces related to instrumentation, such as acoustic/non-acoustic and electronic/non-electronic music, presence of voice, gender of vocals, and dark/bright timbre color (Table 3.5).

Table 3.5: Datasets for other classification tasks

Dataset	Comments	Link to data
AcousticBrainz Instrumentation, timbre & voice	In-house audio collection owned by MTG containing annotations: Acoustic/non-acoustic music (321 full tracks + excerpts). Electronic/non-electronic music (332 full tracks + excerpts). Music with voice/instrumental music (1000 track excerpts, 500 per class). Vocal music gender (male/female) (3311 full tracks). Bright/dark timbre (3000 track excerpts, 1500 per class).	https://acousticbrainz.org/datasets/accuracy

[14] http://www.music-ir.org/mirex/wiki/2015:Music/Speech_Classification_and_Detection

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

4 Conclusion

In this deliverable we have given an introduction to different descriptors for the automatic annotation of music samples and music pieces. We have grouped them in low-level and high-level descriptors, putting out emphasis on providing pointers to existing implementations of some of the referenced descriptors as well as pointers to datasets that can be used for their training and evaluation.

This document should serve as the basis for further research on automatic music annotation and for the final definition of the descriptors to be included in the music annotation tools that will be released within the AudioCommons project (deliverables D4.2, 4.3, 4.7, 4.8, D4.12 and D4.13).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

5 References

[Aljanaki15] Aljanaki A., Wiering F., Veltkamp R. C. (2015). Studying emotion induced by music through a crowdsourcing game. Information Processing and Management, 52(1), 115-128.

[ANSI60] American Standard Acoustical Terminology (American National Standards Institute, New York)

[Aucouturier09] Aucouturier, J. J. (2009). Sounds like teen spirit: Into the grounding of everyday musical terms. Computational insights In J. Minett and W. Wang (Eds.) Language, Evolution and the Brain , Frontiers in Linguistics, 35-64. Taipei: Academia Sinica Press.

[Barrington07] Barrington L., Chan A., Turnbull D., and Lanckriet, G. (2007). Audio Information Retrieval using Semantic Similarity. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, 725–728.

[Benetos11] Benetos, E., and Dixon, S. (2011). Multiple-instrument polyphonic music transcription using a convolutive probabilistic model. In 8th Sound and Music Computing Conference, 19–24.

[Benetos13a] Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., and Klapuri, A. (2013). Automatic music transcription: Challenges and future directions. Intelligent Information. Systems, 41(3), 407–434.

[Benetos13b] Benetos, E., and Dixon, S. (2013). Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. Journal of the Acoustical Society of America (JASA), 133, 1727–1741.

[Benetos13c] Benetos, E., Cherla, S., and Weyde, T. (2013). An efficient shift-invariant model for polyphonic music transcription. In Proc. Int. Workshop Machine Learning and Music, ECML/PKDD.

[Berenzweig03] Berenzweig, A., Ellis, D. P. W., and Lawrence, S. (2003). Anchor space for classification and similarity measurement of music. In International Conference on Multimedia and Expo (ICME), 29–32.

[Bergstra06] Bergstra, J., Casagrande, N., Erhan, D., Eck, D., and Kegl, B. (2006). Aggregate features and ada boost for music classification. Machine Learning, 65(2-3), 473–484.

[Bittner14] Bittner, R., Salamon, J., Tierney, M., Mauch, M., Cannam, C. and Bello, J. (2014). Medleydb: a multitrack dataset for annotation-intensive mir research. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

[Bittner15] Bittner, R., Salamon, J., Essid, S., and Bello, J. (2015). Melody extraction by contour classification. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

[Böck15] Böck, S., Krebs, F., and Widmer, G. (2015). Accurate Tempo Estimation based on Recurrent Neural Networks and Resonating Comb Filters. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

[Boersma93] Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proceedings of the Institute of Phonetic Sciences 17,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

97–110.

[Bogdanov11] Bogdanov, D., Serrà, J., Wack, N., Herrera, P., and Serra, X. (2011). Unifying low-level and high-level music similarity measures. IEEE Transactions on Multimedia, 13(4), 687–701.

[Bogdanov13] Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., and Serra, X. (2013). ESSENTIA: An audio analysis library for music information retrieval. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

[Bogdanov16] Bogdanov, D., Porter A., Herrera P., and Serra X. (2016). Cross-collection evaluation for music classification tasks. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

[Bosch12] Bosch, J. J., Janer, J., Fuhrmann, F., and Herrera, P. (2012). A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals.In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

[Bosch15] Bosch, J. J., and Gómez, E. (2015). Melody extraction by means of a source-filter model and pitch contour characterization. In Music Information Retrieval Evaluation eXchange (MIREX).

[Bosch16a] Bosch, J. J., Bittner, R., Salamon, J., and Gómez, E. (2016). A comparison of melody extraction methods based on source-filter modelling. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

[Bosch16b] Bosch, J. J., Marxer, R., and Gómez, E. (2016). Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music. Journal of New Music Research, DOI: 10.1080/09298215.2016.1182191.

[Brossier07] Brossier, P. M. (2007). Automatic Annotation of Musical Audio for Interactive Applications. PhD Thesis, Queen Mary University of London, London, UK.

[Brown93] Brown, J. C. (1993). Determination of the meter of musical scores by autocorrelation. The Journal of the Acoustical Society of America, 94(4), 1953-1957.

[Camacho08] Camacho, A., and Harris, J. (2008).A sawtooth waveform inspired pitch estimator for speech and music. The Journal of Acoustical Society of America, 124(3), 1638-1652.

[Cancela08] Cancela, P. (2008). Tracking melody in polyphonic audio. In Music Information Retrieval Evaluation eXchange (MIREX).

[Celma06] Celma, O., Herrera, P., and Serra, X. (2006). Bridging the Music Semantic Gap. In European Semantic Web Conference (ESWC) Workshop on Mastering the Gap: From Information Extraction to Semantic Representation.

[Chai01] Chai, W. and Vercoe, B. (2001). Folk music classification using hidden Markov models. In Proceedings of International Conference on Artificial Intelligence.

[Chai05] Chai, W., and Vercoe, B. (2005). Detection of Key Change in Classical Piano Music. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

[Chang11] Chang, C. C., and Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2(3), 27:1–27:27.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

[Chew01] Chew, E. (2001). Modeling tonality: Applications to music cognition. In Proceedings of the 23rd Annual Meeting of the Cognitive Science Society, 206-211.

[Corona15] Corona, H., and O'Mahony, M. P. (2015). An Exploration of Mood Classification in the Million Songs Dataset. In 12th Sound and Music Computing Conference.

[Coutinho14] Coutinho, E., Weninger, F., Schuller, B., & Scherer, K. R. (2014). The Munich LSTM-RNN Approach to the MediaEval 2014 Emotion in Music Task.

[Davies07] Davies, M. E., and Plumbley, M. D. (2007). Context-dependent beat tracking of musical audio. IEEE Transactions on Audio, Speech and Language Processing, 15(3), 1009–1020.

[DeCheveigne02] De Cheveigne, A., and Kawahara, H. (2002). YIN, a Fundamental Frequency Estimator for Speech and Music. The Journal of the Acoustical Society of America, 111(4), 1917–1930.

[Degara12] Degara, N., Rua, E. A., Pena, A., Torres-Guijarro, S., Davies, M. E., and Plumbley, M. D. (2012). Reliability-Informed Beat Tracking of Musical Signals. IEEE Transactions on Audio, Speech, and Language Processing, 20(1), 290–301. doi:10.1109/TASL.2011.2160854

[Dixon01] Dixon, S. (2001). Automatic extraction of tempo and beat from expressive performances. Journal of New Music Research, 30, 39–58. doi:10.1076/jnmr.30.1.39.7119

[Dixon03] Dixon, S., Elias, P., and Gerhard W. (2003). Classification of dance music by periodicity patterns. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

[Downie08] Downie, X. H. J. S., Laurier, C., & Ehmann, M. B. A. F. (2008). The 2007 MIREX audio mood classification task: Lessons learned. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 462-467.

[Dressler12] Dressler, K. (2012). Towards Computational Auditory Scene Analysis: Melody Extraction from Polyphonic Music. In Proceedings of the 9th International Symposium on Computer Music Modelling and Retrieval (CMMR), 319–334.

[Duan10] Duan, Z., Pardo, B., and Zhang, C. (2010). Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. IEEE Transactions on Audio, Speech, and Language Processing, 18(8), 2121–2133.

[Duan14] Duan, Z., Han, J., and Pardo, B. (2014). Multi-pitch streaming of harmonic sound mixtures. IEEE/ACM Transactions on Audio, Speech, and Language Processing. , 22(1), 138–150.

[Durrieu10a] Durrieu, J., Richard, G., David, B., and Févotte, C. (2010). Source/filter model for unsupervised main melody extraction from polyphonic audio signals. IEEE Transactions on Audio, Speech, and Language Processing, 18(3), 564–575.

[Durrieu10b] Durrieu, J. L. (2010). Automatic Extraction of the Main Melody from Polyphonic Music Signals. EDITE Institut TELECOM, TELECOM ParisTech, CNRS LTCI.

[Durrieu11] Durrieu, J. L., David, B., and Richard, G. (2011). A musically motivated mid-level representation for pitch estimation and musical audio source separation. IEEE Journal on Selected Topics on Signal Processing.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

[EBU16] EBU. (2016). Loudness Metering: "EBU MODE" Metering to Supplement EBU R 128 Loudness Normalization.

[Eerola08] Eerola, T., and Ferrer, R. (2008). Instrument library (MUMS) revised. Music Perception: An Interdisciplinary Journal, 25(3), 253-255.

[Eerola09] Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 621-626.

[Eerola11] Eerola, T. and Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. Psychology of Music, 39(1), 18-49.

[El-Maleh00] El-Maleh K., Klein M., Petrucci G., Kabal P. (2000). Speech/music discrimination for multimedia applications. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2445-2448.

[Ellis10] Ellis, D., and Weller, A. (2010). The 2010 LABROSA chord recognition system. In Music Information Retrieval Evaluation eXchange (MIREX).

[Elowsson13] Elowsson, A., Friberg, A., Madison, G., and Paulin, J. (2013). Modelling the Speed of Music using Features from Harmonic/Percussive Separated Audio. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 481–486.

[Emiya10] Emiya, V., Badeau, R., and David, B. (2010). Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. IEEE Transactions on Audio, Speech, Language Processing, 18(6), 1643–1654.

[Eyben16] Eyben, F. (2016). Real-time speech and music classification by large audio feature space extraction. Springer.

[Faraldo16] Faraldo, Á., Gómez, E., Jordà, S., and Herrera, P. (2016). Key Estimation in Electronic Dance Music. In European Conference on Information Retrieval, 335-347.

[Font16] Font, F., and Serra, X. (2016). Tempo Estimation for Music Loops and a Simple Confidence Measure. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

[Fritsch12] Fritsch, J. (2012). High quality musical audio source separation. M.S. thesis, Centre for Digital Music, Queen Mary University of London, London, UK.

[Fu11] Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2011). A survey of audio-based music classification and annotation. IEEE Transactions on Multimedia, 13(2), 303-319.

[Fuentes12] Fuentes, B., Liutkus, A., Badeau, R., and Richard, G. (2012). Probabilistic model for main melody extraction using constant-Q transform. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5357–5360.

[Fuhrmann11] Fuhrmann, F., and Herrera, P. (2011). Quantifying the Relevance of Locally Extracted Information for Musical Instrument Recognition from Entire Pieces of Music. In Proc. of the Int. Conf. on Music Information Retrieval, 239-244.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

[Fuhrmann12] Fuhrmann, F. (2012). Automatic musical instrument recognition from polyphonic music audio signals (Doctoral dissertation, PhD thesis, Universitat Pompeu Fabra).

[Fujishima99] Fujishima, T. (1999). Realtime chord recognition of musical sound: A system using common lisp music. In *Proc. of the Int. Computer Music Conference (ICMC)*, 464-467.

[Giannakopoulos15] Giannakopoulos, T. (2015).pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PloS one* 10 (12), e0144610.

[Gainza09] Gainza, M. (2009). Automatic musical meter detection. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 329–332.

[Gainza11] Gainza, M., & Coyle, E. (2011). Tempo detection using a hybrid multiband approach. IEEE transactions on audio, speech, and language processing, 19(1), 57-68.

[Gerhard03] Gerhard, D. (2003). Pitch extraction and fundamental frequency: History and current techniques. Technical Report, Department of Computer Science, University of Regina.

[Ghodasara15] Ghodasara, V., et al. Speech/music classification using block based MFCC features. In Music Information Retrieval Evaluation eXchange (MIREX)..

[Gkiokas12] Gkiokas, A., Katsouros, V., Carayannis, G., and Stafylakis, T. (2012). Music Tempo Estimation and Beat Tracking By Applying Source Separation and Metrical Relations. In Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 7, 421–424.

[Gomez06] Gómez, E. (2006). Tonal description of music audio signals. Department of Information and Communication Technologies. PhD Thesis

[Gómez09] Emilia Gómez, Martín Haro, and Perfecto Herrera. Music and geography: Content description of musical audio from different parts of the world. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), pages 753–758, 2009.

[Goto02] Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (2002). RWC Music Database: Popular, Classical, and Jazz Music Databases. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 287-288, October 2002.

[Goto03] Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (2003). RWC Music Database: Music genre database and musical instrument sound database. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 229-230.

[Goto99] Goto, M., & Muraoka, Y. (1999). Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions. Speech Communication, 27(3), 311-335.

[Gouyon03] Gouyon, F., & Herrera, P. (2003). Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors. In Audio Engineering Society Convention 114. Audio Engineering Society.

[Gouyon06] Gouyon, F., Klapuri, A., Dixon, S., Alonso, M., Tzanetakis, G., Uhle, C., and Cano, P. (2006). An Experimental Comparison of Audio Tempo Induction Algorithms. in IEEE Transactions on Audio, Speech and Language Processing, 14(5), 1832–1844.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

[Grézl07] Grézl, F., Karafiát, M., Kontár, S., and Cernocky, J. (2007). Probabilistic and bottle-neck features for LVCSR of meetings. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, 757-757.

[Grosche11] Grosche, P., and Müller, M. (2011). Extracting Predominant Local Pulse Information from Music Recordings. IEEE Transactions on Audio, Speech and Language Processing, 19(6), 1688–1701. doi:10.1109/TASL.2010.2096216

[Guaus05] Guaus, E., and Herrera, P. (2005). The rhythm transform: towards a generic rhythm description. Proc. International Computer Music Conference (ICMC), 7–10. http://doi.org/10.2527/jas.2007-0459

[Guaus09] Guaus, E. (2009) Audio content processing for automatic music genre classification: descriptors, databases, and classifiers. PhD thesis, Universitat Pompeu Fabra.

[Hajimolahoseini15] Hajimolahoseini, H., Amirfattahi, R., Soltanian-Zadeh, H., and Gazor, S. (2015). Instantaneous fundamental frequency estimation of non-stationary periodic signals using nonlinear recursive filters. IET Signal Processing, vol. 9, 143–153

[Hall09] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, vol. 11(1).

[Hamel09] Hamel, P., Wood, S. and Eck, D., 2009, October. Automatic Identification of Instrument Classes in Polyphonic and Poly-Instrument Audio. In ISMIR (399-404).

[Han09] Han, B. J., Ho, S., Dannenberg, R. B., & Hwang, E. (2009). Smers: Music emotion recognition using support vector regression.In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 651–656.

[Han16] Han, Y., Kim, J., & Lee, K. (2016). Deep convolutional neural networks for predominant instrument recognition in polyphonic music. arXiv preprint arXiv:1605.09507.

[Harte05] Harte, C., and Sandler, M. (2005). Automatic chord identification using a quantised chromagram. In Audio Engineering Society Convention 118. Audio Engineering Society.

[Herrera06] Herrera-Boyer, P., Klapuri, A., & Davy, M. (2006). Automatic classification of pitched musical instrument sounds. In Signal processing methods for music transcription, 163-200.

[Homburg05] Homburg, H., Mierswa, I., Möller, B., Morik K., and Wurst, M. (2005). A benchmark dataset for audio classification and clustering. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), pages 528–531.

[Hu07] Hu, X., and Downie, J. S. (2007). Exploring Mood Metadata: Relationships with Genre, Artist and Usage Metadata. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 67-72.

[Khadkevich11] Khadkevich, M., and Omologo, M. (2011). Time-frequency reassigned features for automatic chord recognition. in IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 181-184.

[Klapuri06a] Klapuri, A., and Davy, M. (2007). Signal processing methods for music transcription.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Springer Science & Business Media.

[Klapuri06b] Klapuri, A. Multiple fundamental frequency estimation by summing harmonic amplitudes (2016). In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 216–221.

[Klapuri06c] Klapuri, A., Eronen, A., and Astola, J. (2006). Analysis of the meter of musical signals. IEEE Transactions on Audio, Speech, and Language Processing, 14(1), 342–355.

[Korzeniowski16] Korzeniowski F., and Widmer G. (2016), Feature Learning for Chord Recognition: The Deep Chroma Extractor. in Musical Audio. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)

[Krebs13] Krebs, F., Böck, S., & Widmer, G. (2013). Rhythmic Pattern Modeling for Beat and Downbeat Tracking in Musical Audio. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 227-232.

[Kubera14] Kubera, E., Wieczorkowska, A. A., and Skrzypiec, M. (2014). Influence of feature sets on precision, recall, and accuracy of identification of musical instruments in audio recordings. In International Symposium on Methodologies for Intelligent Systems, 204-213. Springer International Publishing.

[Kumaresan14] Kumaresan, R., Peddinti, V. K., & Cariani, P. (2014). Auditory-inspired pitch extraction using a Synchrony Capture Filterbank and phase alignment. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5959-5963.

[Kursa09] Kursa, M., Rudnicki, W., Wieczorkowska, A., Kubera, E. and Kubik-Komar, A. (2009). Musical instruments in random forest. In International Symposium on Methodologies for Intelligent Systems, 281-290. Springer Berlin Heidelberg.

[Laurier09] Laurier, C., Meyers, O., Serra, J., Blech, M., and Herrera, P. (2009). Music mood annotator design and integration. In 7th International Workshop on Content-Based Multimedia Indexing (CBMI'09), 156-161.

[Laurier10] Laurier, C., Meyers, O., Serrà, J., Blech, M., Herrera, P., & Serra, X. (2010). Indexing music by mood: Design and integration of an automatic content-based annotator. Multimedia Tools and Applications, 48(1), 161-184.

[Liu10] Liu, J., and Xie, L. (2010). Svm-based automatic classification of musical instruments. In International Conference on Intelligent Computation Technology and Automation.

[Livshin03] Livshin, A., and Rodet, X. (2003). The importance of cross database evaluation in musical instrument sound classification: A critical approach. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)

[Luo14] Luo, L., Bruck, G. H., and Jung, P. (2014). A Novel Fundamental Frequency Estimator Based on Harmonic Pattern Match for Music Signals. IEEE International Symposium on Multimedia

[Mandel05] Mandel, M. I., and Ellis, D. P. W. (2005). Song-level features and support vector machines for music classification, in International Conference on Music Information Retrieval (ISMIR), 2005, 594–599.

[Mauch10] Mauch, M., and Dixon, S. (2010). Approximate Note Transcription for the Improved

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Identification of Difficult Chords. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 135-140.

[Mauch14] Mauch, M., and Dixon, S. (2014). pYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions, in IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), May 2014, 659–663.

[McKay09] McKay C., and Fujinaga, I. (2009). jMIR: Tools for automatic music classification, in International Computer Music Conference (ICMC), 2009, 65–68.

[Mires13] Serra, X., Magas, M., Benetos, et al. (2013). Roadmap for Music Information Research. G. Peeters, Ed.

[Moore03] Moore, B. C. J. (2003). An Introduction to the Psychology of Hearing, 5th ed. San Diego, CA: Academic Press.

[Müller15] Müller, M. (2015). Fundamentals of Music Processing. Springer.

[Oliveira10] Oliveira, J. L., Gouyon, F., Martins, L. G., and Reis, L. P. (2010). IBT: A real-time tempo and beat tracking system. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 291–296. Retrieved from http://repositorio.ucp.pt/handle/10400.14/4800

[Patel01] Patel, A. D., and Balaban, E. (2001). Human pitch perception is reflected in the timing of stimulus-related cortical activity, Nat. Neurosci. 4, 839–844

[Peeters04] Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. CUIDADO IST Project Report.

[Percival14] Percival, G., and Tzanetakis, G. (2014). Streamlined tempo estimation based on autocorrelation and cross-correlation with pulses. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(12), 1765–1776.

[Poliner05] Poliner, G., and Ellis, D. (2005). A classification approach to melody transcription. In Proc. 6th Int. Conf. Music Information Retrieval, London, 161–166.

[Poliner07a] Poliner, G., and Ellis, D. (2007). A discriminative model for polyphonic piano transcription. EURASIP Journal on Advanced Signal Processing 154–162.

[Poliner07b] Poliner, G. E., Ellis, D. P. W., Ehmann, F., Gómez, E., Steich, S., and Ong, B. (2007) Melody transcription from music audio: Approaches and evaluation. IEEE Trans. Audio, Speech, Language Processing, 15(4), 1247–1256.

[Porter15] Porter, A., Bogdanov, D., Kaye, R., Tsukanov, R., and Serra, X. (2015). Acousticbrainz: a community platform for gathering music information obtained from audio, In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

[Porter16] Porter, A., Bogdanov, D., and Serra, X. (2016). Mining metadata from the web for AcousticBrainz. In International Digital Libraries for Musicology workshop (DLfM).

[Quinton16] Quinton, E., Sandler, M., and Dixon, S. (2016). Estimation of the Reliability of Multiple Rhythm Features Extraction from a Single Descriptor. In Proc. of the Int. Conf. on Acoustics, Speech

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

and Signal Processing (ICASSP), 256–260.

[Roma15] Roma, G. (2015). Algorithms and representations for supporting online music creation with large-scale audio databases. Universitat Pompeu Fabra.

[Russel80] Russel, J. A. (1980). A circumplex model of affect. in Journal of Personal Social Psychology, vol. 39, 1161–1178.

[Salamon12] Salamon, J., & Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6), 1759-1770.

[Salamon14] Salamon, J., Gomez, E., Ellis, D., and Richard, G. (2014). Melody Extraction from Polyphonic Music Signals: Approaches, applications, and challenges. IEEE Signal Process. Mag., 31:118–134.

[Scaringella06] Scaringella, N., Zoia, G., and Mlynek, D. (2006) Automatic genre classification of music content: a survey. Signal Processing Magazine, IEEE, 23(2):133–141.

[Schedl14] Schedl, M., Gómez E., and Urbano J. (2014). Music Information Retrieval: Recent Developments and Applications. Foundations and Trends in Information Retrieval. 8(2-3), 127-261.

[Schindler12] Schindler, A., Mayer, R., and Rauber, A. (2012) Facilitating comprehensive benchmarking experiments on the Million Song Dataset. In International Society for Music Information Retrieval Conference (ISMIR), pages 469–474, 2012.

[Schmidt11] Schmidt, E. M. and Kim, Y. E. (2011). Modeling musical emotion dynamics with conditional random fields. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)

[Schnitzer11] Schnitzer, D., Flexer, A., Schedl, M., and Widmer, G. (2011). Using mutual proximity to improve content-based audio similarity," In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 79–84.

[Schreiber15] Schreiber, H. (2015). Improving genre annotations for the Million Song Dataset, In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

[Seppanen01] Seppanen, J. (2001). Tatum grid analysis of musical signals. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

[Sha'ath11] Sha'ath, I. (2011). Estimation of key in digital music recordings. Master's Thesis, Birkbeck College, University of London, London, UK.

[Shawe-Taylor05] Shawe-Taylor, J. S., and Meng, A. (2005). An investigation of feature models for music genre classification using the support vector classifier. 604-609.

[Sheh03] Sheh, A., and Ellis, D. P. (2003). Chord segmentation and recognition using EM-trained hidden markov models. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 183-189.

[Shepard64] Shepard, R. N. (1964). Circularity in judgments of relative pitch. The Journal of the Acoustical Society of America, 36(12), 2346-2353.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

[Silla08] Silla Jr, C. N., Koerich, A. L., and Kaestner, C. A. (2008). The Latin Music Database, In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 2008, 451–456.

[Slaney02] Slaney, M. (2002). Mixtures of probability experts for audio retrieval and indexing. In Proceedings of the Int. Conf. on Multimedia and Expo (ICME), 345–348.

[Slaney11] Slaney, M. (2011). Web-Scale Multimedia Analysis: Does Content Matter?. IEEE Multimedia, 18(2), 12–15.

[Soleymani13] Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C. Y., and Yang, Y. H. (2013). 1000 songs for emotional analysis of music. In Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia.

[Song13] Song, Y., Dixon, S., Pearce, M. and Halpern, A. (2013). Do Online Social Tags Predict Perceived or Induced Emotional Responses to Music? In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 89–94.

[Sonnenburg10] Sonnenburg, S., Rätsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., De Bona, F., Binder, A., Gehl, C. and Franc, V. (2010) The SHOGUN Machine Learning Toolbox, Journal of Machine Learning Research, 11, pp. 1799–1802.

[Sordo12] Sordo, M. (2012). Semantic Annotation of Music Collections: A Computational Approach, UPF, Barcelona, Spain.

[Steenbergen14] Steenbergen, N. (2014). Chord Recognition with Stacked Denoising Autoencoders. Master Thesis

[Stoter15] Stoter, F.-R., Werner, N., Bayer, S. and Edler, B. (2015). Refining fundamental frequency estimates using time warping. In Proceedings of EUSIPCO 2015, Nice, France, September 2015.

[Sturm12] Sturm, B. L. (2012). An analysis of the GTZAN music genre dataset. In International ACM Workshop on Music Information Retrieval with User-centered and Multi-modal Strategies (MIRUM'12), pages 7–12, 2012.

[Sturm14] Sturm, B. L. (2014). A Simple Method to Determine if a Music Information Retrieval System is a Horse, IEEE Transactions on Multimedia, 16(6), 1636–1644.

[Sturm15] Sturm, B. L. (2015). Faults in the Latin Music Database and with its use. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)

[Su15a] Su, L. and Yang, Y.-H.(2015) Combining spectral and temporal representations for multipitch estimation of polyphonic music, Audio, Speech, and Language Processing, IEEE/ACM Transactions on 23, no. 10, 1600–1612.

[Su15b] Su, L. and Yang, Y.-H. (2015). Escaping from the Abyss of Manual Annotation: New Methodology of Building Polyphonic Datasets for Automatic Music Transcription, in Int. Symp. Computer Music Multidisciplinary Research (CMMR)

[Tzanetakis00] Tzanetakis, G. and Cook, P. (2000). Marsyas: A framework for audio analysis, Organised sound, 4(3), 169–175.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

[Tzanetakis02a] Tzanetakis, G., Essl, G., and Cook, P. (2002). Human Perception and Computer Extraction of Musical Beat Strength. In Proc. of the Int. Conf. on Digital Audio Effects (DAFx), 257–261. doi:10.1.1.20.6271

[Tzanetakis02b] Tzanetakis, G., and Cook, P. (2002). Musical Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing, 10(5):293–302.

[Varewyck13] Varewyck, M., Martens, J. P., and Leman, M. (2013). Musical Meter Classification with Beat Synchronous Acoustic Features, DFT-based Metrical Features and Support Vector Machines. Journal of New Music Research, 42(3), 267–282. doi:10.1080/09298215.2013.798738

[Vickers01] Vickers, E. (2001). Automatic Long-term Loudness and Dynamics Matching. In Proceedings of the 111th Audio Engineering Society Convention. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=9888

[Vincent10] Vincent, E., Bertin, N. and Badeau, R. (2010). Adaptive harmonic spectral decomposition for multiple pitch estimation, IEEE Trans. Audio, Speech, Lang. Process., 18(3), 528–537.

[Wang13] Wang, F. and Zhang, X. (2013). Research on CRFs in music chord recognition algorithm. J. Comput, 8, 1017.

[Williams99] Williams G., Ellis D.P.W. (1999). Speech/music discrimination based on posterior probability features. In Proceedings of the 6th European Conference on Speech Communication and Technology, 687-690.

[Wu14] Wu, F.-H. F. and Jang, J.-S. R. (2014). A Supervised Learning Method for Tempo Estimation of Musical Audio. In Proc. of the Mediterranean Conf. on Control and Automation (MED), 599–604.

[Xu03] Xu, C., Maddage, N. C., Shao, X., Cao, F. and Tian, Q. (2003). Musical genre classification using support vector machines, in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), 2003, 429–432.

[Yeh14] Yeh, C.C.M., Wang, J.C., Yang, Y.H. and Wang, H.M. (2014). Improving music auto-tagging by intra-song instance bagging. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2139-2143.

[Yang08] Yang, Y.-H., Lin, Y.-C., Su, Y.-F. and Chen, H. H. (2008). A regression approach to music emotion recognition. IEEE Trans. Audio, Speech, Lang. Process., 16(2), 448–457.

[Yang14] Yang, N., Ba, H., Cai, W., Demirkol, I. and Heinzelman, W. (2014). BaNa: A Noise Resilient Fundamental Frequency Detection Algorithm for Speech and Music. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(12), 1833-1848.

[Yeh12] Yeh, T.-C., Wu, M.-J., Jang, J.-S., Chang, W.-L. and Liao, I.-B. (2012). A hybrid approach to singing pitch extraction based on trend estimation and hidden Markov models, in IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan, 457–460.

[Zapata11] Zapata, J. R. and Gómez, E. (2011). Comparative Evaluation and Combination of Audio Tempo Estimation Approaches. In Proc. of the AES Int. Conf. on Semantic Audio, 198 – 207.

[Zapata12] Zapata, J. R., Holzapfel, A., Davies, M. E., Oliveira, J. L., and Gouyon, F. (2012). Assigning a

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Confidence Threshold on Automatic Beat Annotation in Large Datasets. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR), 157–162.

[Zapata14] Zapata, J. R., Davies, M. E., and Gómez, E. (2014). Multi-Feature Beat Tracking. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(4), 816–825. doi:10.1109/TASLP.2014.2305252

[Zen07] Zenz, V. (2007). Automatic chord detection in polyphonic audio data. Na. Master Thesis

[Zhang09] Zhang, M.L., Peña, J.M. and Robles, V. (2009). Feature selection for multi-label naive Bayes classification. Information Sciences, 179(19), 3218-3229.

[Zhou15] Zhou, X. and Lerch, A. (2015). Chord detection using deep learning. In Proc. of the Int. Conf. on Music Information Retrieval (ISMIR).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382