



PHENICX

D3.13: Auto-tagger that predicts reliable semantic labels on the granularity of musical segments

| | |
|---|--|
| Grant Agreement nr | 601166 |
| Project title | Performances as Highly Enriched aNd Interactive Concert eXperiences |
| Project acronym | PHENICX |
| Start date of project (dur.) | Feb 1st, 2013 (3 years) |
| Document reference | PHENICX-WD-WP3-JKU-150131-Autotagger-1.1 |
| Report availability | PU - Public |
| Document due Date | Jan 31, 2015 |
| Actual date of delivery | Jan 31, 2015 |
| Leader | JKU |
| Reply to | Markus Schedl (markus.schedl@jku.at) |
| Additional main contributors (authors name / partner acr.) | Hamid Eghbal-zadeh (JKU) Bernhard Lehner (JKU) Ali Nikrang (JKU) Tom Collins (formerly JKU) Jan Schlüter (OFAI) Thomas Grill (OFAI) |
| Document status | First Draft |

Project funded by ICT-7th Framework Program from the European Commission



Table of Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 4 |
| 1.1 | Overview | 4 |
| 1.2 | Main objectives and goals | 4 |
| 1.3 | Methodology | 4 |
| 2 | User Studies on Semantic Concepts | 6 |
| 2.1 | Descriptors from music perception | 6 |
| 2.2 | Free-form descriptors during a concert | 8 |
| 3 | Automatic Segmentation based on Audio | 11 |
| 3.1 | State of the Art | 11 |
| 3.2 | Approach | 11 |
| 3.3 | Evaluation | 12 |
| 4 | Detecting and Visualizing Themes and Motifs from Symbolic Representations | 14 |
| 4.1 | State of the Art | 14 |
| 4.2 | Approach | 14 |
| 4.2.1 | Detection of Themes and Motifs | 14 |
| 4.2.2 | “Pattern Viewer” Visualization | 15 |
| 4.3 | Evaluation | 15 |
| 5 | Surveying Music Using Textural Sound Qualities | 17 |
| 5.1 | State of the Art | 17 |
| 5.2 | Approach | 17 |
| 5.3 | Visualization Example | 18 |
| 6 | Instrument Activity Detection | 19 |
| 6.1 | State of the Art | 19 |
| 6.1.1 | State of the Art in Instrument Activity Detection (IAD) | 19 |
| 6.1.2 | Limitations of the State of the Art | 19 |
| 6.2 | Approach | 20 |
| 6.2.1 | Features | 20 |
| 6.2.2 | Datasets | 20 |
| 6.2.3 | Proposed instrument detection systems | 20 |
| 6.3 | Evaluation | 22 |
| 6.3.1 | Non-classical music | 22 |
| 6.3.2 | Classical music | 22 |
| 7 | Detection of Activity Classes | 24 |
| 7.1 | State of the Art | 24 |
| 7.2 | Approach | 24 |
| 7.3 | Evaluation | 26 |
| 8 | Conclusion | 27 |

EXECUTIVE SUMMARY

The aim of this deliverable **D3.13** is to provide methods that are capable of extracting musically meaningful descriptors from the audio signal of recordings. We approach this task using machine learning techniques to elaborate methods that predict semantic concepts on the granularity of musical segments. The major achievements reported in this deliverable are summarized in the following.

Two **user studies to assess human agreement between semantic concepts** for describing music were conducted. The first was organized as a structured online survey, in which participants had to indicate the perceived strength of emotions, tempo, complexity, and kinds of instruments for each of 15 expert-defined segments of Beethoven's 3rd symphony "Eroica". The second study stimulated short descriptions during a live performance of RCO, in which attendees used their mobile phones to indicate particularly remarkable events in the performance. Both studies evidence that there is little agreement among different music listeners on music descriptions.

A novel **music segmentation** algorithm from audio, based on Convolutional Neural Networks, is proposed. It automatically detects segment boundaries in music recordings, exploiting perceptually informed spectrogram features. Evaluation on a standardized test collection shows that our method is able to substantially outperform the previous state of the art.

In addition to segmentation from the audio, we exploit a recently proposed segmentation algorithm for symbolic representations in order to develop a novel **user interface to explore music by its structural components and by key**. This "PatternViewer" is empirically evaluated in a user study on Beethoven's 1st and 3rd symphonies and turns out to be particularly useful for people less experienced in classical music, who represent an important target group of PHENICX.

Features to describe **textural sound qualities** of music in terms of timbral, temporal, and structural properties of sound on the segment-level are proposed. In addition, we present a new user interface, "Snakeskin", to explore a music piece by these properties and give an example for Beethoven's 3rd symphony.

We approach the problem of **instrument activity detection** and **instrument group activity detection** from audio by proposing novel methods that draw from research in speaker verification. We show on a public corpus of non-orchestra music that our methods produce competitive results, close to the current state-of-the-art algorithms. We further assemble a corpus of 11 different performances of Beethoven's 3rd symphony "Eroica", together with instrument and instrument group annotations, and present first promising results on this corpus.

A method to **detect activity classes**, such as music or applause, was already proposed in **D4.3**. In the deliverable at hand, we propose an extension to this method to deal with the detection of **singing voice**. We evaluate the method on various datasets, among others a **manually annotated corpus of four operas**. This corpus itself represents a valuable asset for future research. We show that our method performs at least as well as other state-of-the-art approaches, while being much less computationally expensive.

The insights gained and the methods developed in this deliverable support, among others, **WP6** "Exploration and Interaction". In particular, **Task 6.1** "Visualisation of music pieces and their performances" benefits from the "PatternViewer" visualization method for musical segments and the "Snakeskin" visualization of sound qualities, which are proposed here. Also **Task 6.2** "Personalised multimodal information system" draws from **D3.13's** methods to identify semantic concepts, in that the personalized system requires a variety of information to select from in order to tailor its content to the user.

1 INTRODUCTION

1.1 Overview

The deliverable **D3.13** responds to **WP3, Task 3.2** of the PHENICX project, as described in the Description of Work (DoW). The goal of this task is to provide methods for multifaceted and musically meaningful analysis of audio streams and performances. To this end, we elaborate techniques employing advanced signal processing and machine learning approaches to address various musical aspects, such as instrumentation, segments/motifs, tonality, or melody. These aspects, which can be regarded as semantic labels, are to be learned on the level of segments of a music piece.

In the deliverable at hand, we focus on the following aspects: First, we perform two **user studies** in order to figure out which semantic concepts listeners agree on, and could hence be robustly learned (Section 2). According to the DoW, the concepts are to be learned on the level of musical segments. We thus need to elaborate methods to **segment music pieces** into somehow meaningful segments that could represent themes or motifs. We approach this task from the audio side (Section 3) and from the MIDI side (Section 4). Based on the identified segments, among others, we elaborate methods that describe the **textural qualities** of sound and music (Section 5). Another category of semantic concepts we address is information about “what is going on” in the piece at a certain moment. To approach this task, we present several approaches to detect **activity of instruments** from the audio signal (Section 6) and to identify **general activity classes**, such as music, speech, applause, silence, and singing voice (Section 7). Eventually, Section 8 summarizes the main contributions of this deliverable and points out some future work.

1.2 Main objectives and goals

The main objective of the deliverable at hand is to provide semantically meaningful descriptors of a musical piece. In particular, these descriptors should not be learned on the level of an entire piece, rather on more fine-grained shorter snippets, or segments, of the piece under consideration. The rationale behind this work is to provide methods that serve the following purposes, among others.

- **Learning:** to learn about the musical structure of a composition, e.g., motifs, themes, key, and modulations;
- **Searching:** to search for particular excerpts or segments in a music piece, which exhibit certain musical or sound qualities;
- **Navigating:** to permit the listener to navigate in a given piece according to various aspects, e.g., “jump to the next passage that sounds similar to the current one” or “jump to the next passage where a solo instrument is playing”;
- **Browsing:** to offer the listener visualizations of structure and sound qualities that can help them exploring a music piece in detail.

1.3 Methodology

Our research is carried out obeying the rules of highest scientific standards, in our opinion. Where possible, we evaluate our methods on existing and publicly available corpora and com-

pare them to state-of-the-art methods. In the case of this deliverable, we additionally had to create some “ground truth” datasets ourselves since suited datasets have not been available before, for instance, for the instrument group activity detection in classical orchestra music and for the singing voice detection in classical operas.

2 USER STUDIES ON SEMANTIC CONCEPTS

Prior to elaborating methods that predict semantic labels from audio, we first had to identify which semantic concepts listeners use to describe music and which concepts they perceive in music. For this purpose, we performed two user studies: one conducted in an orchestra setting, where listeners could mark arbitrary points in time and attach a free-form description to them, and one online survey in which we exposed listeners to predefined segments identified by expert and asked them about their perception according to a variety of dimensions. The former connects to the use case “Capture the Moment” and was carried out as a collaboration between UPF, VD, RCO, and JKU. The latter is an extension to an offline study (carried out previously by UPF and MIT) and was conducted as a joint endeavor of JKU and UPF. Both studies are detailed in the following.

2.1 Descriptors from music perception

In order to assess which semantic descriptors are suited best to learn from audio data, we conducted an online user study. To this end, we used 15 segments of Beethoven’s 3rd symphony, “Eroica”, which were manually identified by music experts. This online study can be regarded as a follow-up to another study, carried out by UPF and MIT, which solely focused on the perception of emotions in classical music and in which 26 music experts participated (with an average of 6.3 years of musical education). Since one main goal of PHENICX is to reach out to new audiences, our follow-up online study was not restricted to people already knowledgeable in classical music. Rather it should include a wide range of people with varying musical expertise.

To achieve this goal, we recruited participants from colleagues, by posting to mailing lists and in social media, and by a mail to all students of JKU. Eventually, more than 200 participants took the survey, among which 178 completed it. Completing the questionnaire took around 40 minutes per participant. We asked participants a range of questions, split into three categories. Screenshots of the survey are provided in Figures 1, 2, and 3; details about the options available to participate for each answer, as well as their numeric coding for the following analysis, are provided in Table 1. First, the participants in the survey had to provide some general personal information (related to demographics and inclination to music and to classical music, in particular), cf. Figure 1. In the next step, they were asked to fill in a personality questionnaire, cf. Figure 2, since we also wanted to investigate relationships between music perception and personality traits. This questionnaire is the standardized “Ten Item Personality Measure” [23]. Afterwards, the participants were presented a carefully designed questionnaire, which they had to fill in for 15 segments of Beethoven’s 3rd symphony, “Eroica”, cf. Figure 3. The segmentation was performed manually by music experts to ensure a maximum of coherence and consistency. The questionnaire shown to participants for each segment was composed of four categories of perceptual aspects: emotions, tempo, complexity, and instrument types. The **emotion** descriptors were taken from the “Geneva Emotion Music Scale” (GEMS) [75] and four basic human emotions from psychological literature. They can be seen in the upper part of Figure 3 (question 1). We further asked participants to indicate the perceived **tempo**, the perceived **complexity**, and the number of **kinds of instruments** of the segment (questions 2, 3, and 4, respectively). Asking for kinds of instruments rather than individual instruments was motivated by the fact that it seemed too hard, even for experts, to identify for instance, whether 2 or 3 flutes are playing. Eventually, in question 5, participants could optionally give an additional description of the segment.

The most important results of the study, concerning the deliverable at hand, are shown in Tables 2 and 3. From the former, we see that participants were slightly biased towards students (minimum age of 18 years, median of 25 years), which does not come as a surprise since we

Basic questionnaire

Please fill-in the following questionnaire. **Note: Fields marked with an asterisk * are mandatory.**

Age*

Gender* Male Female

Country*

How many hours per week do you listen to **classical** music?*

How many hours per week do you listen to **non-classical** music?*

How many hours per week do you play an instrument?*

How many years have you spent studying an instrument (both formal and informal)?*

How many **classical** concerts do you attend per year?*

How many **non-classical** concerts do you attend per year?*

How familiar are you with the classical piece Beethoven's Symphony No. 3 (Eroica)?*

Your Twittername (starting with @)

Your Last.fm ID

Progress: 3/21


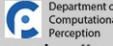



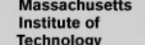











Figure 1: Basic user questionnaire of the survey.

sent a mail to all JKU students, asking them to participate. They also had an additional incentive as RCO thankfully provided us with a stack of CDs to give as a gift to our student participants. As for participants' listening frequency, while there are a few outliers who listen to classical music up to 40 hours per week and to other genres even up to 70 hours, the average participant listens to classical music 3 hours per week and to other genres 11 and a half hours per week. Interestingly, the median for classical music listening (1) is much lower than that of listening to other genres (8). Looking at the distribution of listening frequencies, it seems that participants either love classical music and devote a whole lot of time to it, or do not care at all about it. Less than half of the participants indicated to play an instrument (median of 0), but most had some form of musical education, on average as much as 6 years. Participants attend on average 2 classical and 4 non-classical concerts per year. Compared to the results for listening frequency of classical vs. non-classical music, an interesting observation can be made: lovers of classical music attend many more concerts than listeners of other genres, in relation to the time spent on listening to their preferred genre. Most participants were not or somewhat familiar with Beethoven's "Eroica".

Table 3 shows the agreement among participants for each investigated aspect. Krippendorff's α is used as measure of agreement, and computed for each segment separately (among all user ratings for that segment). The resulting 15 values per aspect are then averaged and shown in Table 3. Unfortunately, as we can see, the agreement for most aspects is very low. Participants do not (0.00–0.20) or at most slightly (0.21–0.40) agree on almost all concepts. The values indicating moderate agreement (0.41–0.60) according to [40] are printed in bold. However, data whose agreement falls below an α of 0.67 is typically discarded [38].

Personality Questionnaire

Please fill-in the following questionnaire. Hover the mouse over the description to get more details.
Note: Fields marked with an asterisk * are mandatory.

| I see myself as ... | Strongly disagree | Moderately disagree | A little disagree | Neither agree nor disagree | A little agree | Moderately agree | Strongly agree |
|-----------------------------------|-----------------------|-----------------------|-----------------------|----------------------------|-----------------------|-----------------------|-----------------------|
| Extraverted, enthusiastic* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Critical, quarrelsome* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Dependable, self-disciplined* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Anxious, easily upset* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Open to new experiences, complex* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Reserved, quiet* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Sympathetic, warm* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Disorganized, careless* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Calm, emotionally stable* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Conventional, uncreative* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Progress: 4/21




Figure 2: Personality questionnaire of the survey.

Table 1: Options available to participants for the questions in the survey, and their numerical encoding for analysis.

| Aspect | Options | Encoding |
|----------------------------|--|----------------------|
| Age | free-form | years |
| Gender | male or female | — |
| Country | list selection from 193 countries | — |
| Listening classical | free-form | hours per week |
| Listening non-classical | free-form | hours per week |
| Playing instrument | free-form | hours per week |
| Musical education | free-form | years |
| Concerts classical | free-form | attendances per year |
| Concerts non-classical | free-form | attendances per year |
| Familiar with "Eroica" | unfamiliar, somewhat familiar, very familiar | 0–2 |
| All personality traits | strongly disagree–strongly agree | 1–7 |
| All emotions | strongly disagree–strongly agree, don't know | 0–6, -1 |
| Perceived tempo | slow, fast, don't know | 0, 1, -1 |
| Perceived complexity | very low–very high, don't know | 0–4, -1 |
| Kinds of instruments | 1, 2, 3, 4, more, don't know | 1, 2, 3, 4, 5, -1 |
| Description of the excerpt | free-form | — |


2.2 Free-form descriptors during a concert

The results of the previous study are well in line with those of another experiment we conducted in the context of our use case "Capture the Moment". To this end, we implemented a very simple prototype application running on smart phones, which allows users to pinpoint and save particular moments in a live performance, just by pressing a button on their device. In addition, listeners could give a short description of the captured moment, as free text. We stored both

Music Tag Questionnaire

Please listen to the music below and answer the questions. Note that the player repeats the current segment, once started. If your environment is not silent, we suggest you to use headphones. Hover the mouse over the description to get more details. **Note: Fields marked with an asterisk * are mandatory.**

Musical piece number 1/15:



1. When I listen to this excerpt, I perceive the music as ...

| description | Strongly disagree | Disagree | Neither agree nor disagree | Agree | Strongly agree | Don't know |
|--------------------|-----------------------|-----------------------|----------------------------|-----------------------|-----------------------|-----------------------|
| Transcendence* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Peacefulness* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Power* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Joyful activation* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Tension* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Sadness* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Anger* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Disgust* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Fear* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Surprise* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Tenderness* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

2. How do you perceive the tempo in this excerpt?

| description | Slow | Fast | Don't know |
|------------------|-----------------------|-----------------------|-----------------------|
| Perceived tempo* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

3. When I listen to this excerpt, I perceive it as complex.


| description | Very low | Low | Medium | High | Very high | Don't know |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Perceived complexity* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

4. How many kinds of instruments do you perceive in this excerpt? (Note: if you perceive 10 violins, this counts as 1 kind of instrument. If you perceive a cello, a flute and a violin, it counts as 3.)
 Hint: Here is a list of instruments that are playing in the orchestra. But some of them might not be playing in this excerpt:
[Flute](#), [Oboe](#), [Clarinet](#), [Bassoon](#), [French horn](#), [Trombone](#), [Timpani](#), [Violin](#), [Viola](#), [Cello](#), [Contrabass](#)

| description | 1 | 2 | 3 | 4 | More | Don't know |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Kinds of instruments* | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

5. Describe this excerpt with a couple of words please. Separate each word with a comma (,).

Description of the excerpt




Progress: 5/21 

Figure 3: Music tag questionnaire of the survey.

the time stamp and the short description in a database. We performed an experiment during an RCO concert, a performance of Shostakovich 5th symphony, which took place on October 11, 2014. Thirty concert goers were equipped with the application and indicated a total of 290 markers (“capturings”) during the performance. A preliminary investigation of the collected data showed that about 50% of the markers were attached a description. When taking a closer look at these descriptions, we found that the vast majority of them can be categorized into three classes: (i) instruments, e.g., “viola”, “celesta”, “harp”, “flute solo”, (ii) indications of liking,

Table 2: Basic statistics of the participants.

| Aspect | μ | σ | med. | min. | max. |
|-------------------------|---------|----------|------|------|------|
| Age | 27.4745 | 8.0355 | 25 | 18 | 67 |
| Listening classical | 3.0561 | 5.4304 | 1 | 0 | 40 |
| Listening non-classical | 11.4311 | 11.4403 | 8 | 0 | 70 |
| Playing instrument | 2.2781 | 5.0673 | 0 | 0 | 40 |
| Musical education | 6.4286 | 5.7717 | 5 | 0 | 28 |
| Concerts classical | 2.3061 | 4.8799 | 1 | 0 | 40 |
| Concerts non-classical | 4.4541 | 7.6576 | 2 | 0 | 70 |
| Familiar with "Eroica" | 0.8469 | 0.6540 | 1 | 0 | 2 |

Table 3: Mean, standard deviation, and agreement for investigated aspects of music perception. Bold face is used to denote moderate agreement. Scores are averaged over all 15 segments.

| Aspect | μ | σ | Krippendorff's α |
|-------------------|--------|----------|-------------------------|
| Transcendence | 2.4207 | 1.0852 | 0.0092 |
| Peacefulness | 3.1517 | 0.8922 | 0.4405 |
| Power | 1.4302 | 0.9940 | 0.4189 |
| Joyful activation | 2.0391 | 1.1581 | 0.3139 |
| Tension | 1.6780 | 1.2309 | 0.2102 |
| Sadness | 1.7247 | 1.1486 | 0.3095 |
| Anger | 0.5251 | 0.7593 | 0.2925 |
| Disgust | 0.4888 | 0.6990 | 0.1216 |
| Fear | 0.8933 | 0.9999 | 0.2617 |
| Surprise | 1.3820 | 1.0996 | 0.0560 |
| Tenderness | 2.8150 | 1.0401 | 0.3455 |
| Tempo | 0.0284 | 0.1666 | 0.4998 |
| Complexity | 2.0347 | 0.7843 | 0.1034 |
| Instruments | 3.9750 | 0.9447 | 0.0743 |

e.g., "great", "nice", "sweet", and (iii) emotions and feelings, e.g., "beautiful", "exciting", "tense", "crazy". While the unstructured free-form nature of the descriptions did not allow for a more detailed quantitative analysis, the empirical observation of the descriptions and time stamps, again, showed that different listeners provided very different kinds of descriptors and used a heterogeneous vocabulary to describe music.

The consistent results of these two studies, i.e., that **semantic descriptors vary too strongly between listeners**, even when they are provided in a structured way of a limited set to choose from, required us to slightly adapt the focus of this deliverable. We hence decided not to focus on a fully fledged "auto-tagger" that is capable of learning arbitrary labels for arbitrary segments in a music piece, as this seems unachievable given the results of our two studies. Instead, we elaborated methods capable of identifying general musical concepts that are as objective as possible. In particular, we developed methods that

- segment an audio stream, i.e. identify meaningful boundaries between musically coherent parts in the piece, and visualize the identified segments, which can well represent motifs or themes,
- assign descriptions of textural sound qualities to short segments of audio and visualize them,
- identify whether an instrument or instrument group is active at a certain moment in time from an audio stream,
- identify certain activity classes from audio, such as music, speech, silence, and applause, with a particular focus on singing voice detection.

3 AUTOMATIC SEGMENTATION BASED ON AUDIO

There are two principal ways to automatically produce a description of a music piece from an audio recording in terms of labeled, characterized segments: (i) run characterization algorithms that continuously monitor the audio data and output instantaneous information, such as the instruments playing, then segment this information into coherent parts, or (ii) run a segmentation algorithm that divides the audio recording into coherent parts, then characterize the segments. For this deliverable, we focused on the second option. We researched ways to segment an audio recording of a music piece into musically coherent parts, aiming to reproduce segmentations manually done by music experts. Our approach turned out to considerably outperform existing methods in terms of agreement with these human annotations.

3.1 State of the Art

An overview paper to audio structure analysis by Paulus et al. [57] distinguishes three fundamental approaches to segmentation: novelty-based (detecting transitions between contrasting parts), homogeneity-based (identifying sections that are consistent with respect to their musical properties), and repetition-based (building on the determination of recurring patterns).

Novelty is typically computed using Self-Similarity Matrices (SSMs) or Self-Distance Matrices (SDMs) with a sliding checkerboard kernel [17], building on engineered audio descriptors like timbre (MFCC features), pitch, chroma vectors, and rhythmic features [56]. Techniques capitalizing on homogeneity use clustering [18] or state-modelling (HMM) approaches [5], or both [46, 45]. Repeating pattern discovery is performed on SSMs or SDMs [47], and often combined with other approaches [55, 51]. Some algorithms combine all three basic approaches [64].

Interestingly, almost all existing algorithms are hand-designed from end to end. To the best of our knowledge, only two methods are partly learning from human annotations: Turnbull et al. [67] compute temporal differences at three time scales over a set of standard audio features including chromagrams, MFCCs, and fluctuation patterns. Training Boosted Decision Stumps to classify the resulting vectors into boundaries and non-boundaries, they achieve significant gains over a hand-crafted boundary detector using the same features. McFee et al. [51] employ Ordinal Linear Discriminant Analysis to learn a linear transform of beat-aligned audio features (including MFCCs and chroma) that minimizes the variance within a human-annotated segment while maximizing the distance across segments. Combined with a repetition feature, their method defined the state of the art in boundary retrieval.

3.2 Approach

From the review of the state of the art, we obtained two insights: (i) The novelty-based approach alone can already be quite successful, and (ii) learning from annotations of human experts seems superior to designing an algorithm completely by hand.

Hence, for the proposed segmentation approach, we elaborate a decent **boundary detection method**, then use advanced machine learning methods to further reduce the amount of manual engineering compared to the state of the art. Specifically, we use a Convolutional Neural Network (CNN) to detect segment boundaries, and we train it directly on **perceptually informed spectrograms** (logarithmic frequency and logarithmic magnitude representations) instead of complex engineered music-specific features.

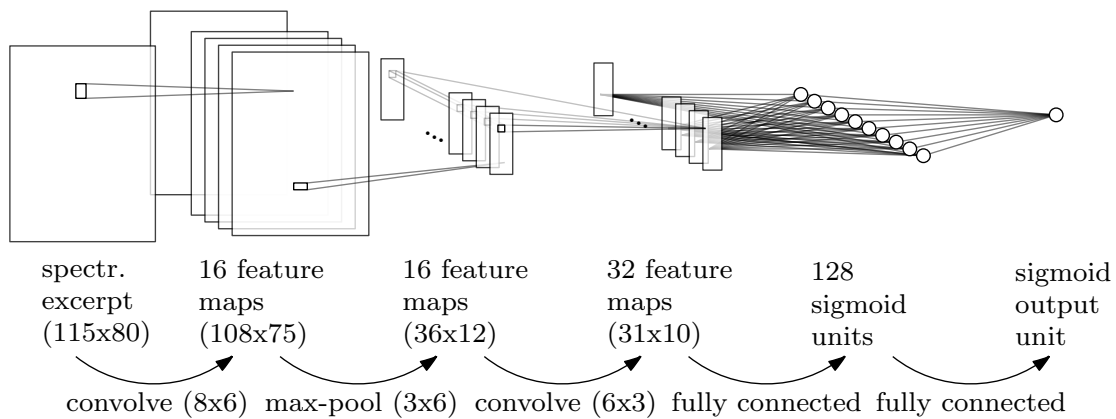


Figure 4: Structure of the Convolution Neural Network we use.

CNNs are **deep neural networks** of a particular architecture that exploits the spatial structure of images, and thus are also suited well for processing spectrograms. We trained a CNN, whose structure is depicted in Figure 4, to classify spectrogram excerpts of about 30 seconds into whether or not they have a boundary near the center. Once trained, we can apply the CNN to highly overlapping excerpts of a new audio recording to obtain a boundary probability curve over time, and report the peaks of that curve as segment boundaries. Initial results were already close to the state of the art, and by developing means to cope with the temporal inaccuracy and scarceness of training examples, we significantly improved over previous methods. For more technical details, please refer to [69].

3.3 Evaluation

We trained our artificial neural network on a set of 633 music recordings, then tested it on a separate set of 487 recordings to estimate how well it generalizes. Since we aimed at developing a general music segmentation method, the music pieces are of different genres and origins, including classical and popular music. Each music piece was annotated by one or two music experts according to a specific set of guidelines in the context of the SALAMI project [65].

An accepted way to measure performance in boundary detection is to compare the predicted boundaries (of an algorithm) to the annotated boundaries (of an expert) and count the number of true positives (prediction is at most X ms from a yet unmatched annotation), false positives (no matching annotation within X ms of prediction) and false negatives (unmatched prediction within X ms of an annotation). From these, we can compute the F-measure, a number between 0.0 and 1.0, balancing how many predicted boundaries are correct (precision) and how many annotated boundaries are correctly found (recall).

In terms of F-measure, our method advances the state of the art from 0.52 to 0.62 for a tolerance of 3,000 ms, and from 0.33 to 0.46 for a tolerance of 500 ms, as shown in Figure 5, in comparison to other recent approaches from literature. The results shown in the figure are all achieved on the same dataset of 487 recordings, mentioned above. **Our method hence does not only detect more of the human-annotated boundaries correctly, it is also significantly more accurate in timing than previous approaches.** To relate these numbers, the F-measure of human experts evaluated against each other is 0.76 for the large tolerance and 0.68 for the smaller one – this is the best any algorithm can hope to achieve. The F-measure of a baseline algorithm that just predicts boundaries at a regular interval is 0.13.

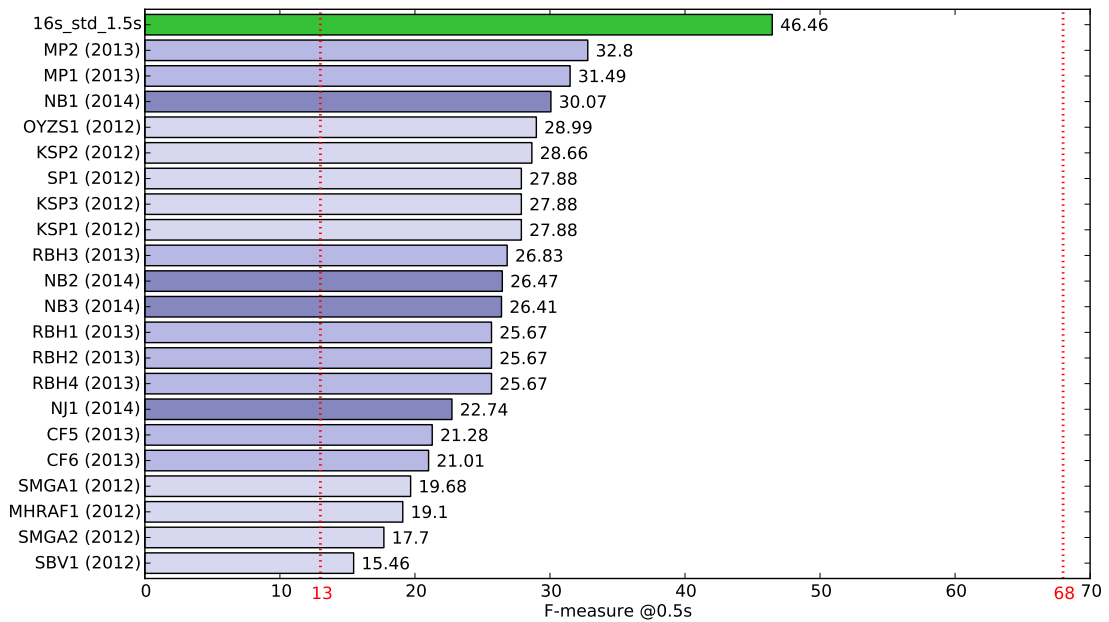


Figure 5: Performance in terms of F-measure achieved by current segmentation algorithms, the topmost bar representing our approach. The vertical dotted lines denote the performance of the baseline and the best achievable performance on the given dataset based on the human annotations.

4 DETECTING AND VISUALIZING THEMES AND MOTIFS FROM SYMBOLIC REPRESENTATIONS

If given a symbolic MIDI representation of the music piece in addition to its audio recording, we are to some extent capable of automatically **identifying musically more meaningful segments**, such as **themes** or **motifs**. The performance of respective algorithms is, however, strongly dependent on the material under consideration; classical orchestra music represents a particularly challenging type of material. Applying an audio-to-score alignment technique, already reported in **D4.3** “Automatic Extraction of Performance-related Parameters from Audio Recordings and Live Performances”, allows users to explore a performance via structural music elements.

The importance of this pattern discovery task is underlined by its recent inclusion in the “Music Information Retrieval Evaluation eXchange” (MIREX) under the name “Discovery of Repeated Themes & Sections”.¹ Organizer of this task is Tom Collins, a former member of the JKU team.

While the algorithm for pattern discovery itself was not developed specifically for PHENICX, but again by JKU's research group [10], we combined it with a key detection algorithm [39] to elaborate an appealing user interface for exploring the structure of classical pieces. More details can be found in [53].

4.1 State of the Art

Geometric representation-based algorithms currently represent the state of the art in music pattern discovery, as evidenced by the novel MIREX task mentioned above. A particularly popular family of these algorithms is the so-called Structure Induction Algorithms (SIA) [52] which aim at finding maximal translatable patterns for a given point set that represents the notes. In short, given the notes in sequential temporal order, the SIA algorithm identifies transpositions of groups of notes. The SIA algorithm has been taken as a basis for later improvements, among others, SIATEC and COSIATEC [52], SIAR [9], SIARCT [11], and SIARCT-CFP [10]. The last one is used in the work at hand. Details can be found in the referenced publications.

4.2 Approach

4.2.1 Detection of Themes and Motifs

The employed **SIARCT-CFP** algorithm [10] extends the SIARCT method by a fingerprinting component to overcome the problem of inexact pattern occurrences, by allowing for matches that are also to a certain degree inexact. The algorithm thus accounts for variability in the data and is motivated by the fact that many themes and motifs occur with some variation, such as transpositions or additional notes. To this end, the output of the SIARCT algorithm is first categorized. Since different occurrences of a certain pattern might be quite similar according to some measure of symbolic music similarity, a threshold is used and only one of such similar occurrences, which is deemed musically most important, is kept. Eventually, a fingerprint of each extracted pattern is computed and matched against a database containing musically meaningful patterns. To this end, the fingerprinting technique developed by Arzt et al. [3] is applied. This procedure accounts for rhythmic variations, transpositions, and time shifts in the patterns under consideration.

¹http://www.music-ir.org/mirex/wiki/2014:Discovery_of_Repeated_Themes_%26_Sections

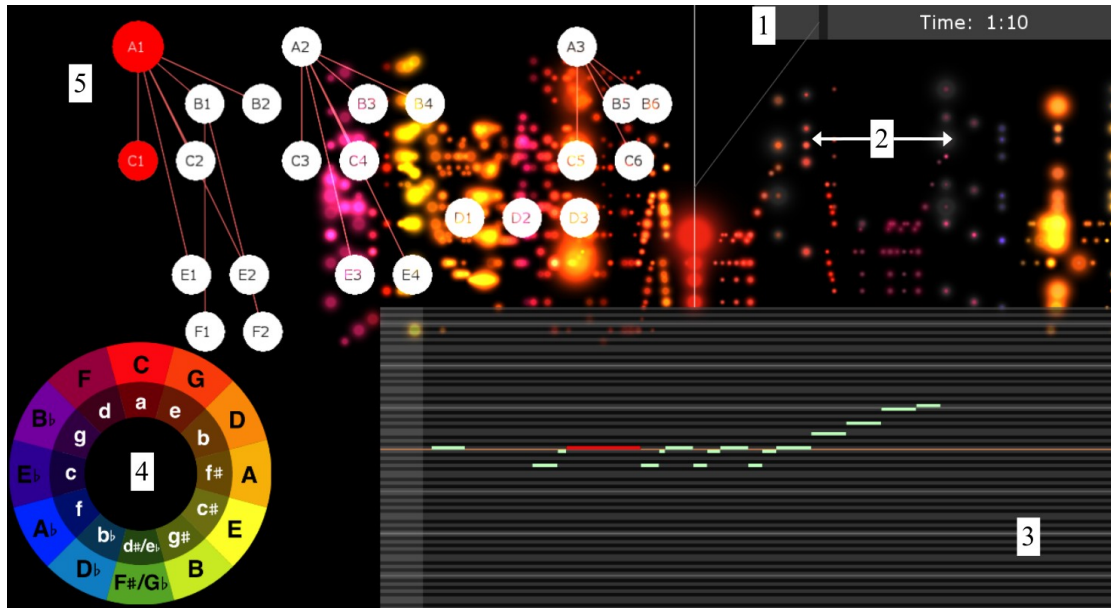


Figure 6: Screenshot of the “PatternViewer” application for Beethoven’s symphony no. 1.

4.2.2 “Pattern Viewer” Visualization

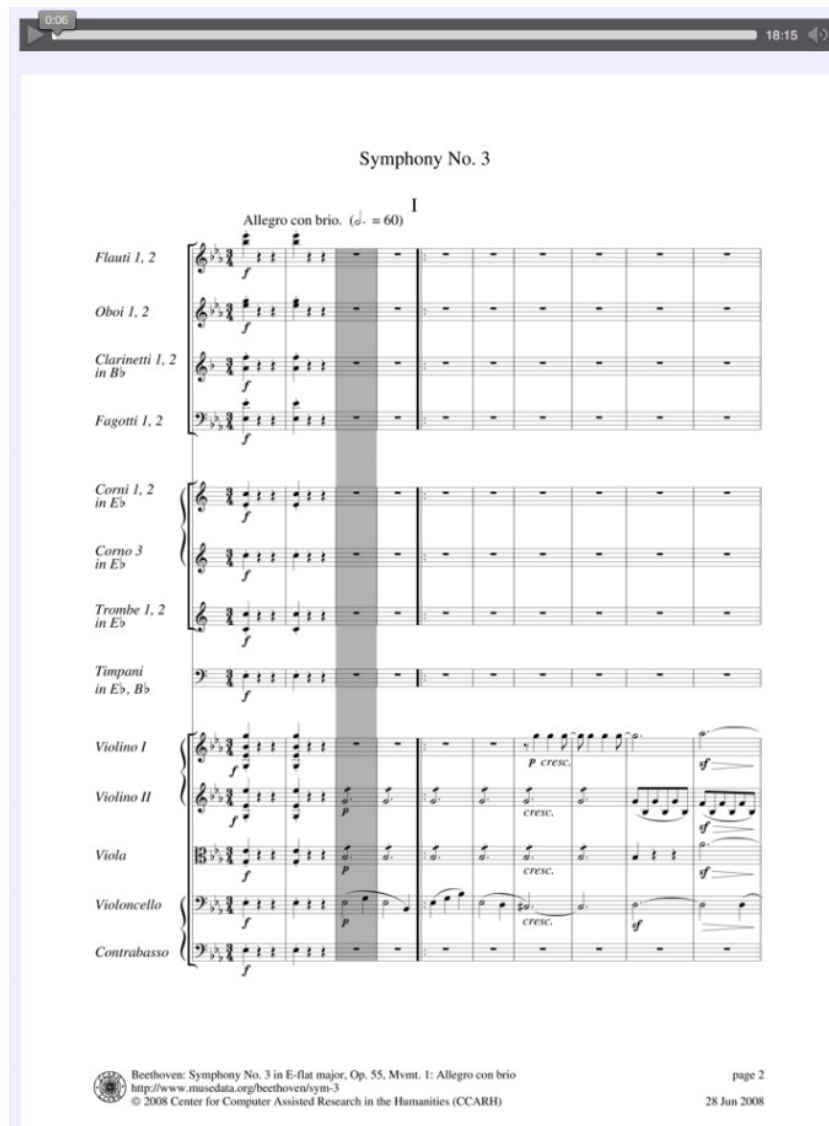
To provide a user-friendly application to explore the identified themes and motifs in a piece, we developed the so-called “**Pattern Viewer**” [54], a user interface that visualizes the structure as well as the key of the piece under consideration. A screenshot is given in Figure 6. Indicator box no. 1 shows the media control bar that can be used to seek within the piece; no. 2 shows a global-scale point-set representation, colored by the current key of the music, with the vertical bar (that can again be dragged) indicating the position within the audio; no. 3 shows a local-scale piano-roll representation, illustrating the contents of a pattern occurrence from the top-left graph; no. 4 shows a colored circle of fifths to identify the current key; no. 5 shows a pendular graph representing the hierarchical, repetitive structure of the current piece, which can be used for jumping to the respective segments. Currently selected is the first occurrence of the main theme C_1 , whose notes are a subset of the larger repeated section A_1 known as the exposition [54].

For the **key estimation** we used the Krumhansl-Schmuckler key finding algorithm applied to a symbolic representation of the piece [39]. The mapping of keys to colors is based on the assumption that tonalities and colors are in some way analogous [58].

For the **discovery of patterns**, we used the SIARCT-CFP algorithm [10], which is capable of accurately identifying themes and motifs in short music pieces. Please note that it is not possible yet to find all the smaller motifs in a symphonic piece automatically. For this kind of material, the SIARCT-CFP algorithm is still able to identify some of the bigger repetitive elements, though. Smaller motifs shown in the pendular graph of Figure 6 are added by hand. The construction of the nodes and edges of the graph is completely algorithmic.

4.3 Evaluation

We evaluated the proposed “PatternViewer” application via an empirical user study [53]. Eighteen subjects (students from JKU) with varying levels of musical expertise interacted with visualizations of two excerpts from Beethoven’s symphonies (no. 1 and no. 3). One visualization



Symphony No. 3

Allegro con brio. (♩. = 60) I

Flauti 1, 2
Oboi 1, 2
Clarinetti 1, 2 in B \flat
Fagotti 1, 2
Corni 1, 2 in E \flat
Corno 3 in E \flat
Trombe 1, 2 in E \flat
Timpani in E \flat , B \flat
Violino I
Violino II
Viola
Violoncello
Contrabasso

Beethoven: Symphony No. 3 in E-flat major, Op. 55, Mvmt. 1: Allegro con brio
<http://www.musedata.org/beethoven/sym-3>
 © 2008 Center for Computer Assisted Research in the Humanities (CCARH)

page 2
28 Jun 2008

Figure 7: Screenshot of the “ScoreViewer” application for Beethoven’s symphony no. 3.

(“ScoreViewer”) showed the staff notation of the music, synchronized automatically to an orchestral recording. This visualization was also developed in the context of the PHENICX project and is described in detail in deliverable **D6.3** “Performance Visualisation Technology”, using the alignment technique reported on in **D4.3** “Automatic Extraction of Performance-related Parameters from Audio Recordings and Live Performances”. A screenshot is provided in Figure 7 to facilitate comparison. The other visualization participants were exposed to was the “PatternViewer” that additionally revealed the music’s repetitive and tonal structure. Classical music appraisal skills of the participants were assessed via multiple-choice questions covering the topics of instrumentation, dynamics, repetition, and tonality. Results indicated that interacting with the “PatternViewer” visualization led to a significant improvement in listeners’ appraisal of the repetitive and tonal structure of a piece of music, compared to interacting with the “ScoreViewer”. The size of this effect was well predicted by the amount of formal musical training, in that **less expert listeners exhibited larger improvements using the “PatternViewer” than more experienced listeners**. With the current high level of interest in applications to improve a user’s linguistic, numeric, or musical prowess, these findings for music appraisal skills are of more general significance.

5 SURVEYING MUSIC USING TEXTURAL SOUND QUALITIES

This section is concerned with the characteristics of textural sounds and their application for surveying music. While most music visualizations are event-based, i.e. focused on notes or motifs, we rather want to take a look at perceptually relevant timbral and micro-temporal qualities that are descriptive for the overall sonic appearance of the music. For a survey of a musical piece on a larger time scale, temporal details can be neglected and perceived qualities be regarded as quasi-stationary, that is, being significant for durations of at least a few seconds. This leads us to the notion of a **sound texture** where acoustic qualities are more or less stationary [63].

5.1 State of the Art

The overwhelming bulk of the literature dealing with audio features is about “song” characterization and classification [19], where two of the predominant tasks are genre classification [4] and emotion respectively mood classification [35]. They are usually tackled using a set of audio descriptors combined with machine learning algorithms to unearth potential relationships between feature combinations and the target classes. In most cases, the classification is performed on discrete classes, either genre classes like ‘rock’, ‘pop’, ‘jazz’, ‘classical music’, ‘world music’, etc., or mood classes like ‘happy’, ‘sad’, ‘dramatic’, ‘mysterious’, ‘passionate’, etc.

While the genre concept is not applicable to general – especially textural – sound, the task of modeling emotion respectively affect in sound and music is somewhat comparable to the task of modeling perceptual qualities. With the existence of a continuous representation of emotion in the *valence–arousal* plane [62, 6], Yang et al. [74] formulate music emotion recognition (MER) as a regression problem to predict such arousal and valence values. They test both linear regression and Support Vector Regression (SVR) based on a selection of 18 – mostly spectral – musical features. Alternative approaches have been published in [43, 48, 29], among others.

The prevailing visual representations of musical sound are based on low-level physical parameters, such as instantaneous amplitude or spectral coefficients in the omnipresent waveform or spectrogram displays [1]. However, such visualizations are highly abstract, lacking an intuitive relationship to perceptual attributes of sound.

We have directed our attention to the translation of **humanly accessible higher-level semantic** respectively **affective concepts** into adequate graphical representations. When connecting the auditory to the visual domain, principles of synesthesia come into play. These have been widely explored by Whitney [72], Levin [44] and many others, often putting emphasis on aesthetic aspects and emancipating it as a genre of its own, known as “Visual Music”. For an overview, please refer to [8].

5.2 Approach

In our original research, we followed a three-stage approach: Firstly, we **identified perceptual qualities** relevant for textural sounds. As described in detail in [28], by conducting mixed qualitative-quantitative interviews within the repertory grid framework we elicited 10 bi-polar qualities based on a corpus of 100 textural sounds.

Secondly, in [26] we took the **five most prominent of those qualities**, namely *high–low*, *ordered–chaotic*, *smooth–coarse*, *tonal–noisy*, and *homogeneous–heterogeneous*, covering **timbral**, **temporal**, and **structural properties** of sound, and constructed audio descriptors capable

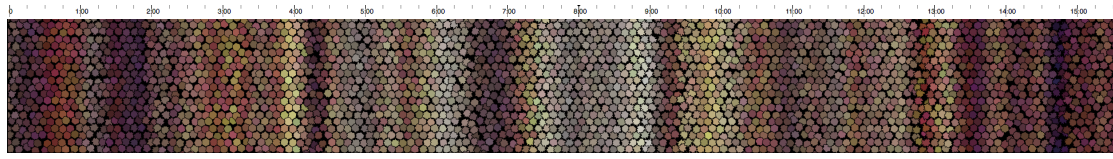


Figure 8: “Snakeskin” visualization of Beethoven’s 3rd symphony, 2nd movement (Adagio).

of modeling those. Finally, in [27] we developed an adequate, **intuitively accessible visualization strategy**, encoding the same five textural qualities. The implementation makes use of tiled maps, essentially combining low-dimensional projection and iconic representation.

The construct *high–low* is encoded in brightness and coloring, *tonal–noisy* in color saturation, *smooth–coarse* in the smoothness of the tile outline, *ordered–chaotic* in the regularity of tile positioning on a grid, and *homogeneous–heterogeneous* in the variance of random deviations of the above parameters.

5.3 Visualization Example

The original intention was to use this kind of visualization as a basis for two-dimensional screen-based sound browsing. In the context of PHENICX, the graphical representation has rather been used in a time-linear fashion, depicting the development of sound qualities along the duration of a piece of music.

As there is no objective “ground truth” for the sound textures at hand, rather than performing a quantitative evaluation, we discuss an exemplary visualization of Beethoven’s 3rd symphony “Eroica”. Figure 8 depicts the visualization of the second movement, performed by RCO under the conduction of Ivàn Fischer. Here, the time resolution is quite coarse, depicting the whole piece of 15:30 minutes at once, revealing its global structure. The time scale can be chosen according to the intended level of detail, bounded by the minimum time support necessary for the calculation of the audio features of about five seconds.

Looking at the figure, we note the *dark* double-bass-carried beginning, becoming more *colorful* within the first minute with the entry of woodwinds. After another double-bass section just before minute 2, we can see some graphical *disorder*, connected to irregular tempo respectively agogics. The smooth and colorful passage around minute 3 is generated by soloistic appearances of woodwinds (bassoon and clarinet). After an alternation of high and low sections around minute 4, we enter a longer section of *roughness* and less saturated colors, indicating more pronounced contours (e.g., staccato notes) and a more complex spectrum. The fugue, starting at around minute 7 is easily distinguishable by its high *order* and *homogeneity*, as well as *rough, grayish* appearance due to the dominance of brass and string instruments. After calming down at minute 9, we find another more energetic section until minute 10, featuring low strings and brass winds, before returning to more *smoothness*, produced by woodwinds. After a more *heterogeneous* section which is terminated by *low* and *rough* soloistic strings just before minute 13, we enter a smooth and dark section, leading to a particularly agogic passage after minute 14, before the movement settles to a calm, contained closing.

6 INSTRUMENT ACTIVITY DETECTION

Instrument activity detection (IAD) is the automatic detection of musical instruments to determine which instrument or instruments are active in a recorded or live audio music excerpt at a certain point in time. In this section, we present our investigations of and methods to IAD for classical music in the context of the PHENICX project.

We first review the state of the art in IAD, then investigate different IAD tasks for classical and non-classical music. As an extension to state-of-the-art approaches, we propose a new method to **detect major instrument groups in polyphonic classical pieces**.

6.1 State of the Art

6.1.1 State of the Art in Instrument Activity Detection (IAD)

The task at hand aims at detecting major active instruments in a classical music piece. Although the focus of PHENICX is classical music, some of the existing methods developed for non-classical music may be generalized and used within the project. Because the problem of accurately identifying all musical instruments solely from the audio signal is too hard, we can re-define the problem as follows:

1. Problem no. 1: Provide labels for monophonic recordings (instrument recognition for solo musical pieces)
2. Problem no. 2: Provide indexes for locating the predominant instrument in a musical mixture (predominant instrument recognition for polyphonic musical pieces)
3. Problem no. 3: Provide indexes for locating all active instruments in a musical mixture (instrument detection for polyphonic musical pieces)

Some work has already been carried out to target the first problem on non-classical music [14, 2, 15, 12]. For the second problem, two main approaches are used on non-classical music: (i) methods that use source separation and (ii) methods that do not use source separation. Promising approaches that use source separation include [7, 71]. Methods of the latter category (no source separation) include [21, 20]. The last and hardest problem (no. 3) can also be approached either with or without source separation techniques [66, 73, 36]. In [16], to simplify the hardest problem no. 3, instead of using perceptual descriptors and detecting the instrument names, a taxonomic classification is done and instrument groups are detected.

Some works, for example Kitahara et al. [37], propose several techniques to improve instrument recognition in duo and trio music. More recent works deal with instrument recognition in polyphonic music. For instance, Fuhrmann et al. [20] propose a method for automatic recognition of predominant instruments with Support Vector Machine (SVM) classifiers trained on features extracted from audio signals. Tzanetakis [68] focus on the detection of voice, while Essid et al. [16] present an approach using a taxonomy-based hierarchical classification, in which the classifiers are trained on combinations of instruments.

6.1.2 Limitations of the State of the Art

Instrument activity detection in polyphonic music is a big challenge. Although some methods have been proposed for the first and second problems, the third problem still needs further investigations. In particular when it comes to classical music, all state-of-the-art methods to IAD are very limited. The three most important reasons are: (i) the lack of an annotated dataset to be used specifically for IAD in classical music (ii) the fact that classical music, especially

orchestral music the PHENICX project focuses on, is per definition highly polyphonic and shows sophisticated forms of instrumental interplay and (iii) for a new IAD task, a new complex set of different features is usually needed.

To address the first point, we prepare a dataset using 11 different performances of Beethoven's 3rd symphony "Eroica". To go one step further towards the second issue, we study the detection of most dominant instrument groups on the provided dataset. To target the third issue, we propose an utterance-level transformation technique which creates a low-dimensional space with a better representation of our instrument classes. We only use MFCC features and a simple k-nearest neighbors (KNN) classifier with $k=1$ and cosine distance. To show that our features also work with other classifiers, a Probabilistic Linear Discriminant Analysis (LDA) is used to model the class distributions.

6.2 Approach

In the context of this deliverable, we elaborate methods that target problems no. 2 and 3. In particular, we focus on the two tasks of (i) predominant IAD and (ii) instrument group activity detection.

6.2.1 Features

Mel-Frequency Cepstrum Coefficients (MFCCs) are features that have proven useful for many audio and music processing tasks [14, 30]. MFCCs provide a compact representation of the spectral envelope and are probably more musically meaningful than other common representations. Even though there are better representations based on MFCCs such as [42], we stay away from feature engineering and focus on the modeling techniques. Additional studies on feature extraction will be conducted later on to further improve results. For the experiments at hand, we have extracted standard MFCCs.

6.2.2 Datasets

Two datasets have been used in our experiments. The first dataset, **IRMAS** [7], is intended for training/testing methods for the automatic recognition of predominant instruments in musical audio, though not specifically classical music. The instruments considered are: cello, clarinet, flute, acoustic guitar, electric guitar, organ, piano, saxophone, trumpet, violin, and human singing voice.

The second dataset was prepared specifically for this task using 11 different performances of Beethoven's 3rd symphony "Eroica". We extracted more than 4,000 excerpts of three seconds length from the recorded CD audio tracks of live performances. More information about the audio files can be found in Table 4. The most dominant instrument groups for each excerpt were computed using an audio-to-score alignment algorithm [22] and music scores for the annotations. Randomly selected alignments were checked manually to validate the correctness of the annotations. The available instrument groups in this **Eroica** dataset are: string, brass, woodwind, and percussive. Because of the infrequent presence of some of these instruments, we decided to distill a set of most frequently co-appearing groups instead of separate classes. These are: string, string-brass-woodwind, and string-woodwind.

6.2.3 Proposed instrument detection systems

Three new instrument detection systems, all based on MFCC features, are proposed to address problem no. 2, hence to identify the predominant instrument. In systems 1 and 2, we use a feature transformation technique as a post-processing step after extracting a set of statistics,

Table 4: Orchestra, conductor, year of performance, and movement(s) used of Beethoven's symphony no. 3 in the Eroica dataset.

| Orchestra | Conductor | Year | Mov. |
|---|-------------|------|---------|
| Berlin Philharmonic Orchestra | Furtwängler | 1952 | 1 |
| NBC Symphony Orchestra | Toscanini | 1953 | 1 |
| Philharmonia Orchestra | Klemperer | 1959 | 1 |
| Berlin Philharmonic Orchestra | Von Karajan | 1963 | 1 |
| Chicaco Symphony Orchestra | Solti | 1973 | 1 |
| Wiener Philharmoniker | Bernstein | 1978 | 1 |
| Chamber Orchestra of Europe | Harnoncourt | 1991 | 1 |
| Orchestre Révolutionnaire et Romantique | Gardiner | 1993 | 1 |
| Tonhalle Orchester Zürich | Zinman | 1998 | 1 |
| London Symphony Orchestra | Haitink | 2005 | 1 |
| Royal Concertgebouw Orchestra | Fischer | 2013 | 1,2,3,4 |

which is called “statistical supervectors” or Baum-Welch statistics [34]. In system 1, a simple KNN classifier is fed with the transformed features. In system 2, we employ Probabilistic Linear Discriminant Analysis (PLDA), instead of KNN to show that our approach is not limited to a single classifier. In system 3, Principal Component Analysis (PCA) is used as a similar transformation approach as i-vectors in systems 1 and 2.

System 1: Ivector-LDA-CosineKNN system This method employs a post-processing method called “i-vector extraction” [13] on the computed MFCCs, which converts features of each audio segment into an information-rich low-dimensional fixed-length-vector. A 1024-component-Gaussian Mixture Model (GMM) is trained for the Universal Background Model (UBM). UBM is a Gaussian mixture model (GMM) which is trained as an instrument-independent model using all training data. From each audio file, a set of statistics are extracted as a supervector. These statistics are the posterior probabilities of the Gaussian components for observations of different instruments. Eventually, a simple KNN classifier with $k=1$ and cosine distance is used to classify the samples. A block diagram of the proposed system is shown in Figure 9.

System 2: Ivector-LDA-PLDA system This system is similar to system 1, but uses a Probabilistic Linear Discriminant Analysis (PLDA) classifier [33, 49], instead of KNN.

System 3: PCA-LDA-PLDA system This system is the same as system 2, except for its use of Principal Components Analysis (PCA) instead of i-vector extraction.

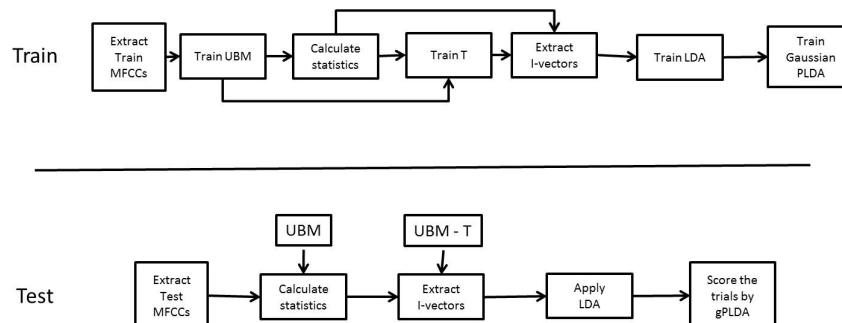


Figure 9: Block diagram of an i-vector based system.

6.3 Evaluation

The IRMAS dataset comprises a training set and 3 test sets. We used this predefined split in our experiments. For the Eroica dataset, we randomly selected 70% of data in each class as our training set and the rest for testing. Because of the lack of annotated data, we only used the training excerpts from “Eroica”, yet using data from other classical pieces will be helpful in the future to achieve more generalization. We calculated performance measures, such as precision, recall, and F-measure separately for each class and also report unweighted averages of the measures, which is proposed in [7] as macro measures. A random baseline and a state-of-the-art method by Bosch et al. [7] is used to compare our results to. The state-of-the-art method uses different sets of features, a feature selection step and a Support Vector Machine (SVM) classifier with a polynomial kernel of degree 4.

6.3.1 Non-classical music

Results achieved on the IRMAS dataset by systems 1, 2, and 3 are shown in Tables 6, 7 and 8, respectively. A comparison of the averaged results between different systems and baselines is given in Table 5. System 1 outperforms the other proposed systems with an average precision of 42.16%, recall of 28.51% and F-measure of 31.86% for the 11 class predominant instrument recognition task. The proposed systems are considerably better than the random baseline, which has an average precision of 15.29%, recall of 8.82%, and F-measure of 9.57%. In comparison with the state-of-the-art method, System 1 has a 3.61 percentage points better recall than the state-of-the-art algorithms. But the state-of-the-art method outperforms our best system in precision and F-measure by 15.64 and 3.04 percentage points, respectively. Because more detailed results for the state-of-the-art method are not provided in [7], a class-wise comparison is not possible; yet by looking at Table 6, we can observe that clarinet (“cla”) is the hardest class, while voice (“voi”), electric guitar (“gel”), and violin (“vio”) are the easiest to detect.

Table 5: Overall results for instrument recognition using different methods, on the **IRMAS** dataset.

| | Prec | Rec | F-measure |
|------|--------------|--------------|--------------|
| SOA | 57.80 | 24.90 | 34.90 |
| RND | 15.29 | 8.82 | 9.57 |
| Sys1 | 42.16 | 28.51 | 31.86 |
| Sys2 | 41.83 | 28.48 | 30.88 |
| Sys3 | 39.62 | 25.26 | 27.31 |

Table 6: Instrument recognition measures (in %) using the **lvector-LDA-CosineKNN** system (no. 1) on the **IRMAS** dataset.

| | cel | cla | flu | gac | gel | org | pia | sax | tru | vio | voi | avg |
|-----------|-------|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Prec | 20.91 | 3.76 | 29.19 | 45.39 | 76.20 | 21.23 | 67.93 | 42.92 | 32.26 | 38.14 | 85.82 | 42.16 |
| Rec | 20.72 | 8.06 | 33.13 | 25.79 | 30.25 | 17.17 | 23.42 | 29.75 | 47.90 | 42.65 | 34.77 | 28.51 |
| F-measure | 20.81 | 5.13 | 31.03 | 32.90 | 43.31 | 18.99 | 34.83 | 35.14 | 38.55 | 40.27 | 49.49 | 31.86 |

Table 7: Instrument recognition measures (in %) using the **lvector-LDA-PLDA** system (no. 2) on the **IRMAS** dataset.

| | cel | cla | flu | gac | gel | org | pia | sax | tru | vio | voi | avg |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Prec | 19.83 | 4.94 | 25.31 | 42.39 | 79.18 | 22.49 | 70.41 | 43.50 | 31.08 | 34.77 | 86.27 | 41.83 |
| Rec | 21.62 | 12.90 | 37.42 | 26.54 | 26.65 | 18.01 | 20.80 | 26.69 | 46.71 | 42.18 | 33.72 | 28.48 |
| F-measure | 20.69 | 7.14 | 30.20 | 32.64 | 39.87 | 20.00 | 32.12 | 33.08 | 37.32 | 38.12 | 48.48 | 30.88 |

Table 8: Instrument recognition measures (in %) using **PCA-LDA-PLDA** system (no. 3) on the **IRMAS** dataset.

| | cel | cla | flu | gac | gel | org | pia | sax | tru | vio | voi | avg |
|-----------|-------|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Prec | 6.13 | 3.42 | 38.69 | 44.27 | 81.46 | 26.68 | 66.90 | 33.94 | 22.35 | 34.77 | 77.19 | 39.62 |
| Rec | 17.12 | 6.45 | 32.52 | 20.93 | 28.45 | 32.96 | 19.30 | 11.35 | 35.33 | 49.76 | 23.66 | 25.26 |
| F-measure | 9.03 | 4.47 | 35.33 | 28.43 | 42.17 | 29.49 | 29.95 | 17.01 | 27.38 | 40.94 | 36.22 | 27.31 |

Table 9: Instrument recognition measures (in %) using the **random baseline** on the **IRMAS** dataset.

| | cel | cla | flu | gac | gel | org | pia | sax | tru | vio | voi | avg |
|-----------|-------|------|-------|-------|-------|------|-------|-------|------|------|-------|-------|
| Prec | 4.76 | 1.21 | 7.66 | 20.33 | 36.26 | 9.96 | 31.95 | 13.45 | 4.87 | 4.94 | 32.73 | 15.29 |
| Rec | 11.71 | 4.84 | 11.66 | 9.16 | 10.51 | 7.48 | 7.74 | 11.35 | 7.78 | 6.16 | 8.62 | 8.82 |
| F-measure | 6.77 | 1.94 | 9.25 | 12.63 | 16.30 | 8.54 | 12.46 | 12.31 | 5.99 | 5.49 | 13.65 | 9.57 |

Table 10: Instrument recognition measures (in %) using the **lvector-LDA-CosineKNN** system (no. 1) on the **Eroica** dataset.

| | string | string-woods | string-brass-woods | avg |
|-----------|--------------|--------------|--------------------|-------|
| Prec | 77.24 | 62.84 | 76.12 | 72.07 |
| Rec | 36.36 | 17.99 | 20.52 | 24.96 |
| F-measure | 49.45 | 27.97 | 32.33 | 36.58 |

6.3.2 Classical music

Since classical music is more challenging than other kinds of music, instead of predicting individual instrument classes, we opted to classify instrument groups. Three major instrument groups have been identified in the Eroica dataset: (i) strings, (ii) woodwinds, and (iii) brass. Since most of the time different instrument groups are active together, we created a new class set by merging those instrument groups that co-occur frequently. The resulting three new classes are (i) strings, (ii) strings-woodwinds (iii) strings-brass-woodwinds. Table 10 shows the results of system 1 on the Eroica dataset, as this system outperformed on both the IRMAS and the Eroica datasets. System 1 achieved an unweighted averaged precision of 72.07%, recall of 24.96%, and F-measure of 36.58% in the three-class problem on the Eroica dataset, which are much higher results than those achieved on the IRMAS set, because only the 3 merged classes are considered.

In summary, our methods achieved an F-measure very close to the state-of-the-art IAD method on the IRMAS dataset, while using much simpler features (only MFCCs) and a much less complex classifier. As for classical orchestra music, we prepared a new dataset “Eroica” for instrument detection and recognition purposes. Our approaches already achieved considerable classification results for the merged instrument groups in this Eroica dataset.

7 DETECTION OF ACTIVITY CLASSES

As a final category of semantic concepts, we target information about the current (musical or non-musical) activity in an audio stream, in particular a recording of a live performance. Methods to determine some of these classes, **music**, **speech**, **applause**, and **silence**, have already been developed as part of deliverable **D4.3** “Automatic Extraction of Performance-related Parameters from Audio Recordings and Live Performances”, elaborated on in Section 5 “Detecting relevant activity classes”. To this end, two new features, the Continuous Frequency Activation (CFA) and Curved Frequency Trajectory (CFT) were adapted and the problem was treated as a binary classification task, which means that all of the four classes could be detected independently of each other. On a manually annotated dataset assembled from various sources, we already achieved quite good results (>97% F-measure for applause and music detection, >93% for speech detection, and >78% for detecting silence). For more details, please refer to **D4.3** and to Pieringer [59].

For the deliverable at hand, we developed, in the same vein as done for **D4.3**, a method to classify a given audio snippet according to whether it contains **singing voice** or not. This is a particularly challenging task as voice is frequently confused with certain kinds of instruments, e.g., some playing styles of electric guitars. Our method is described in detail in [42] and summarized in the following.

7.1 State of the Art

Quite a lot of work has been devoted to the task of identifying singing voice in an audio signal. Recent approaches include Regnier and Peeters [61], whose method is based on thresholding the outputs of vibrato and tremolo detectors. Hsu and Jang [31] use Gaussian Mixture models (GMMs) as states in a fully connected Hidden Markov Model (HMM) and the Viterbi algorithm. Hsu et al. [32] extend [31] by performing harmonic/percussive source separation prior to the actual singing voice detection. Ramona et al. [60] use a very diverse feature set and a Support Vector Machine (SVM) as classifier. The method by Mauch et al. [50], that seems to yield one of the best results achieved so far, employs a complex set of expensive features, based among others on features inferred from the F0 trajectory of the predominant source, e.g., pitch fluctuation or normalized amplitude of harmonic partials. They use HMMs and SVMs as classifier. A more decent literature review can be found in our paper [41]. In summary, all approaches to singing voice detection available so far involve rather complex and computationally expensive features and/or classifiers.

7.2 Approach

In contrast to the previously sketched methods, the approach we elaborated for the deliverable at hand uses a rather **lightweight feature set that is computationally inexpensive** and can thus theoretically also be used for on-line detection of voice in music. We already proposed in [41] a simple method for singing voice detection, which solely uses optimized long-term MFCCs along with their first derivative. It does not require preprocessing and employs a simple median filter to smooth the predictions of the Random Forest classifier used in the actual classification step. Motivated by the remarkably good performance of this very simple method in comparison to much more complex approaches such as [60, 50], we extended our method, focusing on the problem of frequent false positives our previous method [41] was prone to.

In contrast to other state-of-the-art methods, such as [60, 50], our current method, presented in [42], does not presume that the voice is the predominant source in the mix. The method is

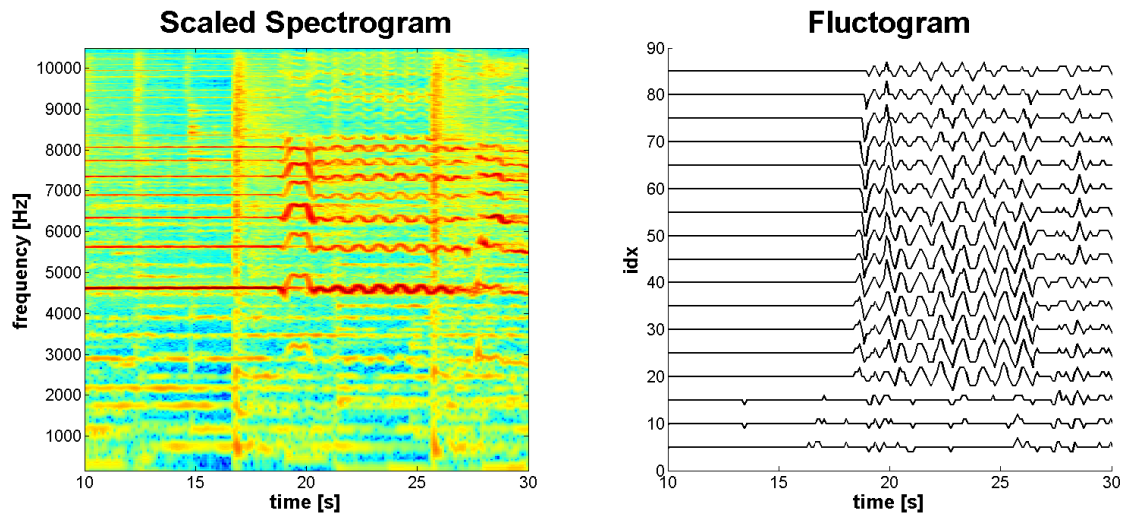


Figure 10: Spectrogram and Fluctogram representation of an audio signal.

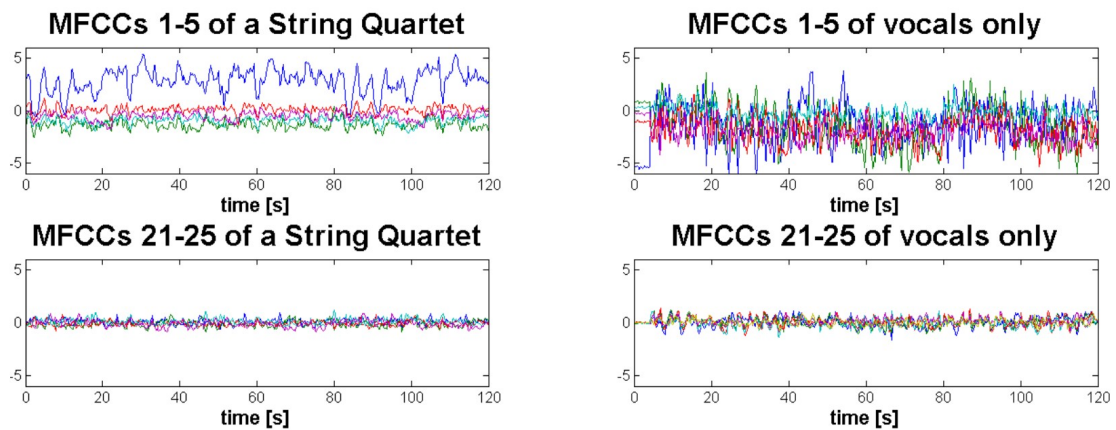


Figure 11: Comparison of the discriminative power of lower and higher MFCCs. The upper two plots depict only the 5 lowest MFCCs for a string quartet and a vocal piece, while the lower two plots show the highest 5 MFCCs for the same two pieces.

based on three features that describe temporal characteristics of the audio signal. The three features are the Fluctogram, the Spectral Contraction, and the Vocal Variance. The **Fluctogram** captures sub-semitone fluctuations of partials by a 17-dimensional feature vector derived from the spectrogram (cf. Figure 10). The **Spectral Contraction** measures how much of the energy in the signal is located in the center of the frequency spectrum. It is simply defined as the ratio between a weighted version of the spectrum and the original spectrum. The **Vocal Variance** feature is inspired by the fact that slow variations of the spectrum are related to changes of the shape of the vocal tract. Since the lowest MFCCs capture such slow variations, we use MFCCs 1–5 as well as their variances over windows of 11 successive frames. An illustration of the usefulness of this feature is given in Figure 11, which reveals that the lowest 5 MFCCs are much better able to reveal differences between instruments (strings in this case) and singing voice than the higher MFCCs. The three proposed features are extended by the Spectral Flatness [25] feature that estimates the noise in an audio signal. The combined feature vector is then fed into a Random Forest classifier. For more details, the reader is kindly referred to [42].

Table 11: Performance measures (in %) of our method in comparison with Ramona et al.’s [60], Mauch et al.’s [50], and Vembu and Baumann’s [70] approaches. Best results for each performance measure are printed in bold face.

| | Ramona et al. [60] | Mauch et al. [50] | Vembu and Baumann [70] | Our approach [42] |
|----------------|--------------------|-------------------|------------------------|-------------------|
| Jamendo | | | | |
| Precision | — | — | 70.8 | 88.0 |
| Recall | — | — | 84.2 | 86.2 |
| F-measure | 84.3 | — | 76.9 | 87.1 |
| Accuracy | 82.2 | — | 77.4 | 88.2 |
| RWC | | | | |
| Precision | — | 88.7 | 82.7 | 87.5 |
| Recall | — | 92.1 | 80.8 | 92.6 |
| F-measure | — | 90.4 | 81.8 | 90.0 |
| Accuracy | — | 87.2 | 81.3 | 87.5 |

7.3 Evaluation

We evaluated our approach using two publicly available corpora that provide singing voice annotations: the **Jamendo** corpus of 93 royalty-free songs annotated by Ramona et al. [60] and the **RWC** music database of 100 songs released by Goto et al. [24] and annotated by Mauch et al. [50]. Since Ramona et al. only report results of their method on the Jamendo set and so do Goto et al. on the RWC set, we could unfortunately solely compare the results achieved by our method on the respective datasets for which figures of merit are available. We further include results reported by Vembu and Baumann in [70] as they are available for both datasets. A summary of the results is given in Table 11. We can thus conclude that **our method, while exploiting rather simple features and being computationally quite inexpensive, is nevertheless at least on par with other state-of-the-art methods.**

In addition to the evaluation experiments carried out in our paper [42], we further evaluated the approach on classical material, in particular, a manually annotated dataset comprising four operas: Don Giovanni and Zauberflöte by Mozart, Madama Butterfly by Puccini, and La Traviata by Verdi. The annotations were not only made on the level of singing voice vs. no singing voice; additionally, the gender of the singer was identified. This dataset thus also represents a valuable resource for future research beyond this deliverable. In a four-fold cross validation setting (using three operas for training and the fourth for testing), our approach achieved an average F-measure of 92.4% and accuracy of 90.9%. We hence conclude that the approach is, in general, **suited for classical music**, even though opera music is quite different in terms of singing style from pop and rock music.

8 CONCLUSION

In this deliverable report, we presented our work towards the identification, extraction, and visualization of semantic concepts from music recordings, in particular on the segment-level. The deliverable's main contributions can be summarized as follows:

- two **user studies** carried out on different user groups (mainly students through an online survey and attendees of a classical concert) to identify which semantic concepts listeners agree on; both studies evidence little agreement among listeners for the vast majority of concepts,
- a new audio music segmentation algorithm based on deep neural networks, which significantly improves over the previous state of the art,
- a novel user interface to explore a music piece by its **themes** and **motifs** and by **key**,
- identification and extraction of features that describe **textural sound qualities** in terms of timbral, temporal, and structural dimensions as well as a visualization tool that illustrates these sound qualities for a given recording,
- novel approaches to **instrument activity detection** and **instrument group activity detection**, which adopt methods used in speaker identification,
- a novel method to **singing voice detection**, which is computationally less expensive, but of equal predictive performance than the current state of the art,
- new **annotated datasets**, in particular, an instrument-annotated collection of 11 different performances of Beethoven's 3rd symphony "Eroica" and a singing voice-annotated corpus of 4 operas (annotations are also available for the gender of the singer), which represent valuable resources for future research,

The methods developed and insights gained in this deliverable will be further exploited in **WP6** "Exploration and Interaction". In particular, the developed "PatternViewer" visualization method for musical segments fits well into **Task 6.1** "Visualisation of music pieces and their performances". In addition, **Task 6.2** "Personalised multimodal information system" requires the methods developed in the deliverable at hand, to identify semantic concepts, as the personalized music information system needs to select from a variety of information about the music piece, to address the specific needs of individual users with different backgrounds.

We foresee several next steps as follow-ups to this deliverable: (i) a deeper analysis of the interrelationship between user characteristics (demographics, education, music experience, personality traits, etc.) and music perception/characterization, (ii) extending our symbolic pattern discovery algorithm to symphony-length classical repertoire, (iii) improving our methods on instrument activity detection and instrument group activity detection, (iv) integrating our methods for singing voice detection with previous methods to identify other activity classes (silence, music, applause, speech), and investigate the identification of joint activities (e.g., music and singing), and (v) extending our annotated datasets for instrument activity and singing voice detection.

Addressing (i), UPF and JKU are currently taking a deeper look into the results of their user studies. To achieve (iii), we will strengthen the collaboration between JKU, OFAI, and TUD and draw from the findings of **D3.10** and **D3.11** to investigate truly multi-modal approaches (using MIDI, audio, image, and video). Drawing from **D4.3** and **D3.13**, we will further strengthen the collaboration between OFAI and JKU to target (iv).

References

- [1] N. Adams. *Analytical Methods of Electroacoustic Music*, chapter Visualization of Musical Signals, pages 13–28. Routledge, 2006.
- [2] G. Agostini, M. Longari, and E. Pollastri. Musical instrument timbres classification with spectral features. *EURASIP Journal on Applied Signal Processing*, 2003:5–14, 2003.
- [3] A. Arzt, S. Böck, and G. Widmer. Fast Identification of Piece and Score Position via Symbolic Fingerprinting. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, Porto, Portugal, October 8-12 2012.
- [4] J.-J. Aucouturier and F. Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.
- [5] J.-J. Aucouturier and M. Sandler. Segmentation of musical signals using hidden markov models. In *Proc. AES 110th Convention*, May 2001.
- [6] L. F. Barrett. Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus. *Cognition and Emotion*, 12(4):579–599, 1998.
- [7] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera. A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In *ISMIR*, pages 559–564, 2012.
- [8] K. Brougher and J. Zilcher. *Visual music: Synaesthesia in art and music since 1900*. Museum of Contemporary Art, 2005.
- [9] T. Collins. *Improved Methods for Pattern Discovery in Music, with Applications in Automated Stylistic Composition*. PhD thesis, Faculty of Mathematics, Computing and Technology, The Open University, UK, 2011.
- [10] T. Collins, A. Arzt, S. Flossmann, and G. Widmer. SIARCT-CFP: Improving Precision and the Discovery of Inexact Musical Patterns in Point-set Representations. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil, November 2013.
- [11] T. Collins, J. Thurlow, R. Laney, A. Willis, and P. H. Garthwaite. A Comparative Evaluation of Algorithms for Discovering Translational Patterns in Baroque Keyboard Works. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, Utrecht, the Netherlands, August 2010.
- [12] B. David and G. Richard. Efficient musical instrument recognition on solo performance music using basic features. In *Audio Engineering Society Conference: 25th International Conference: Metadata for Audio*. Audio Engineering Society, 2004.
- [13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798, 2011.
- [14] A. Eronen. Comparison of features for musical instrument recognition. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 19–22. IEEE, 2001.
- [15] S. Essid, G. Richard, and B. David. *Musical instrument recognition on solo performances*. 2004.
- [16] S. Essid, G. Richard, and B. David. Instrument recognition in polyphonic music based on automatic taxonomies. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):68–80, 2006.

- [17] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 1, pages 452–455 vol.1, 2000.
- [18] J. T. Foote and M. L. Cooper. Media segmentation using self-similarity decomposition. In *Proc. of The SPIE Storage and Retrieval for Multimedia Databases*, volume 5021, pages 167–175, San Jose, California, USA, January 2003.
- [19] Z. Fu, G. Lu, K. M. Ting, and D. Zhang. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2):303–319, April 2011.
- [20] F. Fuhrmann, M. Haro, and P. Herrera. Scalability, generality and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music. In *ISMIR*, pages 321–326. Citeseer, 2009.
- [21] F. Fuhrmann and P. Herrera. Polyphonic instrument recognition for exploring semantic similarities in music. In *Proc. of 13th Int. Conference on Digital Audio Effects DAFx10*, pages 1–8, 2010.
- [22] M. G. M. Gasser, A. Arzt, and G. Widmer. Automatic alignment of music performances with structural differences. 2013.
- [23] S. D. Gosling, P. J. Rentfrow, and W. B. S. Jr. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6):504–528, December 2003.
- [24] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. "RWC music database: Popular, classical, and jazz music databases". In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, volume 2, pages 287–288, 2002.
- [25] J. Gray, A. and J. Markel. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 22(3):207–217, Jun 1974.
- [26] T. Grill. Constructing high-level perceptual audio descriptors for textural sounds. In *Proceedings of the 9th Sound and Music Computing Conference (SMC 2012)*, pages 486–493, Copenhagen, Denmark, 2012.
- [27] T. Grill and A. Flexer. Visualization of perceptual qualities in textural sounds. In *Proceedings of the International Computer Music Conference (ICMC 2012)*, pages 589–596, Ljubljana, Slovenia, 2012.
- [28] T. Grill, A. Flexer, and S. Cunningham. Identification of perceptual qualities in textural sounds using the repertory grid method. In *Proceedings of the 6th Audio Mostly Conference, AM '11*, pages 67–74, New York, NY, USA, 2011. ACM.
- [29] B. Han, S. Ho, R. B. Dannenberg, and E. Hwang. SMERS: Music Emotion Recognition Using Support Vector Regression. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009)*, pages 651–656, October 2009.
- [30] P. Herrera-Boyer, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1):3–21, 2003.
- [31] C.-L. Hsu and J.-S. R. Jang. "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset". *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):310–319, 2010.
- [32] C.-L. Hsu, D. Wang, J.-S. R. Jang, and K. Hu. "A Tandem Algorithm for Singing Pitch Extraction and Voice Separation From Music Accompaniment". *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1482–1491, 2012.
- [33] S. Ioffe. Probabilistic linear discriminant analysis. In *Computer Vision—ECCV 2006*, pages 531–542. Springer, 2006.

- [34] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. A study of interspeaker variability in speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(5):980–988, 2008.
- [35] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. State of the art report: Music emotion recognition: A state of the art review. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 255–266, 2010.
- [36] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Musical instrument recognizer “instrogram” and its application to music retrieval based on instrumentation similarity. In *Multimedia, 2006. ISM’06. Eighth IEEE International Symposium on*, pages 265–274. IEEE, 2006.
- [37] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps. *EURASIP Journal on Applied Signal Processing*, 2007(1):155–155, 2007.
- [38] K. Krippendorff. *Content Analysis – An Introduction to Its Methodology*. SAGE, 3rd edition, 2013.
- [39] C. L. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, New York, USA, 1990.
- [40] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.
- [41] B. Lehner, R. Sonnleitner, and G. Widmer. Towards Light-weight, Real-time-capable Singing Voice Detection. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil, November 2013.
- [42] B. Lehner, G. Widmer, and R. Sonnleitner. On the reduction of false positives in singing voice detection. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
- [43] M. Leman, V. Vermeulen, L. D. Voogdt, D. Moelants, and M. Lesaffre. Prediction of musical affect using a combination of acoustic structural cues. *Journal of New Music Research*, 34(1):39–67, June 2005.
- [44] G. Levin. Painterly interfaces for audiovisual performance. Master’s thesis, MIT Media Arts and Sciences, September 2000.
- [45] M. Levy and M. Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):318–326, Feb 2008.
- [46] B. Logan and S. Chu. Music summarization using key phrases. In *In Proc. IEEE ICASSP*, pages 749–752, 2000.
- [47] L. Lu, M. Wang, and H.-J. Zhang. Repeating pattern discovery and structure analysis from acoustic music data. In *MIR ’04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 275–282, New York, NY, USA, 2004. ACM.
- [48] K. F. MacDorman and S. O. C. Ho. Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *Journal of New Music Research*, 36(4):281–299, 2007.
- [49] P. Matejka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky. Full-covariance ubm and heavy-tailed plda in i-vector speaker verification. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4828–4831. IEEE, 2011.

- [50] M. Mauch, H. Fujihara, K. Yoshii, and M. Goto. "Timbre and Melody Features for the Recognition of Vocal Activity and Instrumental Solos in Polyphonic Music". In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, pages 233–238, 2011.
- [51] B. McFee and D. Ellis. Learning to segment songs with ordinal linear discriminant analysis. In *International conference on acoustics, speech and signal processing*, ICASSP, 2014.
- [52] D. Meredith. Point-set algorithms for pattern discovery and pattern matching in music. In T. Crawford and R. C. Veltkamp, editors, *Content-Based Retrieval*, number 06171 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2006. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.
- [53] A. Nikrang. Interactive Visualisation of Musical Structure with a Focus on Automatic Pattern Discovery. Master's thesis, Johannes Kepler University, Linz, Austria, 2014.
- [54] A. Nikrang, T. Collins, and G. Widmer. PatternViewer: An Application for Exploring Repetitive and Tonal Structure. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, Taipei, Taiwan, October 2014.
- [55] J. Paulus and A. Klapuri. Music structure analysis by finding repeated parts. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, AMCMM '06, pages 59–68, New York, NY, USA, 2006. ACM.
- [56] J. Paulus and A. Klapuri. Acoustic features for music piece structure analysis. In *Conference: 11th International Conference on Digital Audio Effects (Espoo, Finland)*, 2008.
- [57] J. Paulus and A. Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *Trans. Audio, Speech and Lang. Proc.*, 17:12, 2009.
- [58] K. Peacock. Synesthetic perception: Alexander scriabins color hearing. *Journal of Music Perception*, 2(4):483–506, 1985.
- [59] J. Pieringer. On-line Event Detection in Music Performances. Master's thesis, Johannes Kepler University, Linz, Austria, 2014.
- [60] M. Ramona, G. Richard, and B. David. "Vocal detection in music with support vector machines". In *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008*, pages 1885–1888. IEEE, 2008.
- [61] L. Regnier and G. Peeters. "Singing voice detection in music tracks using direct voice vibrato detection". In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009*, pages 1685–1688. IEEE, 2009.
- [62] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, December 1980.
- [63] N. Saint-Arnaud. Classification of sound textures. Master's thesis, MIT Media Lab, Cambridge, MA, USA, September 1995.
- [64] J. Serra, M. Müller, P. Grosche, and J. L. Arcos. Unsupervised detection of music boundaries by time series structure features. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1613–1619. Association for the Advancement of Artificial Intelligence, 2012.
- [65] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and J. S. Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 555–560, 2011.
- [66] H. Sundar, R. HG, and T. Sreenivas. Student's-t mixture model based multi-instrument recognition in polyphonic music. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 216–220. IEEE, 2013.

- [67] D. Turnbull and G. Lanckriet. A supervised approach for detecting boundaries in music using difference features and boosting. In *In Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)*, pages 42–49, 2007.
- [68] G. Tzanetakis. Song-specific bootstrapping of singing voice structure. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, volume 3, pages 2027–2030. IEEE, 2004.
- [69] K. Ullrich, J. Schlüter, and T. Grill. Boundary Detection in Music Structure Analysis using Convolutional Neural Networks. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, Taipei, Taiwan, 2014.
- [70] S. Vembu and S. Baumann. "Separation of vocals from polyphonic audio recordings". In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, London, UK, September 11–15 2005.
- [71] J. J. B. Vicente. Synergies between musical source separation and instrument recognition. *Pro Gradututkielma, Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Espanja*, 2011.
- [72] J. Whitney. *Digital Harmony: On the Complementarity of Music and Visual Art*. McGraw-Hill, Peterborough, N.H., 1980.
- [73] J. Wu, E. Vincent, S. A. Raczynski, T. Nishimoto, N. Ono, and S. Sagayama. Polyphonic pitch estimation and instrument identification by joint modeling of sustained and attack sounds. *Selected Topics in Signal Processing, IEEE Journal of*, 5(6):1124–1132, 2011.
- [74] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):448–457, February 2008.
- [75] M. Zentner, D. Grandjean, and K. R. Scherer. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, 8(4):494–521, August 2008.