D3.4 Methods to compute music content descriptors

<table>
<thead>
<tr>
<th>Grant Agreement nr</th>
<th>601166</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project title</td>
<td>Performances as Highly Enriched aNd Interactive Concert eXperiences</td>
</tr>
<tr>
<td>Project acronym</td>
<td>PHENICX</td>
</tr>
<tr>
<td>Start date of project (dur.)</td>
<td>Feb 1st, 2013 (3 years)</td>
</tr>
<tr>
<td>Document reference</td>
<td>PHENICX-D-WP3-UPF-140710-D3.4MusicContentDescriptors</td>
</tr>
<tr>
<td>Report availability</td>
<td>PU - Public</td>
</tr>
<tr>
<td>Document due Date</td>
<td>August 1st 2014</td>
</tr>
<tr>
<td>Actual date of delivery</td>
<td></td>
</tr>
<tr>
<td>Leader</td>
<td>UPF</td>
</tr>
<tr>
<td>Reply to</td>
<td>Emilia Gómez (UPF) (<a href="mailto:emilia.gomez@upf.edu">emilia.gomez@upf.edu</a>)</td>
</tr>
<tr>
<td>Additional main contributors (authors name / partner acr.)</td>
<td>Markus Schedl (JKU) Agustín Martorell (UPF) Juan J. Bosch (UPF)</td>
</tr>
<tr>
<td>Document status</td>
<td>Final Version (reviewed by JKU)</td>
</tr>
</tbody>
</table>

Project funded by ICT-7th Framework Program from the European Commission
# Table of Contents

1 **Background** ........................................ 4

2 **Introduction** ........................................ 5
   2.1 Main objectives and goals ........................................ 5
   2.2 Methodology ........................................ 5
   2.3 Terminology ........................................ 5
   2.4 Convention ........................................ 6

3 **Onset and note location** ........................................ 7
   3.1 State of the art ........................................ 7
   3.2 Proposed approach ........................................ 7
   3.3 Evaluation ........................................ 8

4 **Multi-pitch and predominant melodic line estimation** ........................................ 9
   4.1 Introduction ........................................ 9
   4.2 Evaluation dataset: definition and annotation ........................................ 9
   4.3 State of the art and proposed approach ........................................ 10
   4.4 Evaluation strategy ........................................ 11
   4.5 Results ........................................ 12
   4.6 Discussion and future work ........................................ 13

5 **Structural and tonal analysis** ........................................ 14
   5.1 Introduction ........................................ 14
      5.1.1 Tonality and structure ........................................ 14
      5.1.2 Orchestration and structure ........................................ 14
      5.1.3 Case study ........................................ 15
   5.2 State of the art ........................................ 15
      5.2.1 Tonal description ........................................ 15
      5.2.2 Structural analysis ........................................ 15
      5.2.3 Instrumentation and orchestration ........................................ 16
   5.3 Methodology ........................................ 16
      5.3.1 Structural ground truth ........................................ 16
      5.3.2 Key estimation ........................................ 16
      5.3.3 Cadence finding ........................................ 17
      5.3.4 Instrumentation ........................................ 18
      5.3.5 Orchestration ........................................ 18
   5.4 Conclusions and future work ........................................ 19

6 **Conclusion** ........................................ 20

7 **List of authors** ........................................ 21
EXECUTIVE SUMMARY

This deliverable presents PHENICX approach for the automatic description of music pieces. Most of the literature in this respect focuses on mainstream popular music, and it has been optimized for this particular repertoire.

The PHENICX project focuses on classical music, in particular on large ensemble settings and symphonic repertoire. This music material is especially challenging due to its complex instrumentation and musical structures. In addition, there is the lack of ground truth data for comparative evaluation.

In this deliverable we first characterize the specificities of the target repertoire with respect to automatic music description. By specific we mean the musical characteristics and derived signal properties which are unique to the symphonic music. We then gather ground-truth annotations from human subjects, in order to be able to evaluate state-of-the-art music content descriptors algorithms and assess the challenges in this particular repertoire. After that, we adapt and design new techniques adapted the PHENICX context. We finally evaluate the improvement with respect to the state of the art.

Music pieces should be described in terms of harmony, tonality, melody and structure and according to different abstraction levels (from signal descriptors to semantic labels) and temporal scopes (descriptors might be related to the whole piece or to a given segment). We include both numerical descriptors and semantic labels. We have focused our research on the following description areas, as they have been identified to be relevant for the defined use cases (see user case document D2.2). These are: (1) location of note events (onset detection), as presented in Section 3; (b) estimation of the predominant pitch and melodic line, covered in Section 4; (c) semantic musical features: we focus on the concepts of structure, orchestration and tonality and we enhance our descriptors with information obtained from the symbolic domain, as explained in Section 5.

This deliverable then proposes and evaluates a comprehensive set of descriptors characterizing the musical piece under consideration, valid for visualization (WP6), retrieval and processing in the context of PHENICX prototypes.
1 BACKGROUND

According to the Description of Work (DoW), this deliverable belongs to WP3: Multimodal musical piece analysis, a work package intended to provide technologies for the automatic description of the music pieces included in the PHENICX repertoire: classical music in large ensemble settings.

In particular, this deliverable addresses Task 3.2: Multifaceted and musically meaningful analysis of audio streams and recordings, leaded by UPF and with contributions by JKU. The goal of this task is to provide technologies for the analysis of the audio material collected in Task 3.1: Audio-visual data acquisition in order to gather musically meaningful descriptors of it. Those descriptors should be related to harmony, tonality, melody and structure and will be related to different abstraction levels and temporal scopes (instantaneous and on the level of musical segments). They include both numerical descriptors and semantic labels.

This deliverable is also partially linked to Task 3.4: Multimodal support for audio processing techniques and to score alignments techniques developed in WP4-Task 4.3: Methods for synchronizing recorded performances with scores or other performances, and for reliable live performance tracking, due to the fact that we exploit the link between audio and score information for musical descriptions. Finally, some descriptors of the audio signal (e.g., pitch, note locations and timbre) are both exploited for characterizing the musical piece (WP3) and the musical performance (WP4), so there are strong links with D4.3 Automatic Extraction of Performance-related Parameters from Audio Recordings and Live Performances as these techniques are contrasted and shared.

Finally, this deliverable is a crucial input for other tasks, in particular for music visualization (WP6-Task 6.1: Visualization of music pieces and their performances), instrument emphasis (WP6-Task 6.3: Acoustic rendering of augmented music performance) and integration into prototypes and analysis of tests results (WP7).
2 INTRODUCTION

2.1 Main objectives and goals

This document addresses the extraction of musically meaningful descriptors from music audio signals. We focus our developments and evaluation in classical music and large ensemble settings, the target repertoire of the PHENICX project. State-of-the-art technologies present a limitation on this particular repertoire that we have addressed within the project.

The target descriptors are related to harmony, tonality, melody and structure and apply to different abstraction levels and temporal scopes. They include both numerical descriptors and semantic labels. We structure the deliverable in the following description areas:

- **Low-level signal feature extraction**: We focus on two main problems, which are crucial in music transcription and representation and valid for both performance and piece characterization:
  - Location of note events (onset detection), as presented in Section 3.
  - Estimation of the main pitches present in the audio signal: multi-pitch estimation, covered in Section 4.
- **Mid-level signal descriptors**: We stress the relevance of detecting predominant melodic lines, covered in Section 4.
- **High-level musical features**: We focus on the concepts of structure, orchestration and tonality. In order to do so, we enrich our audio descriptors with information obtained from the symbolic domain and generate a set of semantic musical descriptors. This aspect is covered in Section 5.

2.2 Methodology

Within this work and over the different tasks, we have adopted the following methodology:

- **Music collection building**: selection of relevant pieces and fragments.
- **Ground truth gathering**: collection of manual descriptors / annotations by human subjects for the considered music collection.
- **Evaluation of state-of-the-art** techniques for automatic music description on the considered dataset, by comparing manual with automatic descriptors using standard evaluation measures.
- **Adaptation** of existing algorithms and **design** of novel and specific approaches.
- **Evaluation** of selected approaches and improvements with respect to the state of the art.
- **Integration** into prototypes and link with related research topics from the PHENICX project, as stated in Section 1.

2.3 Terminology

We use a consistent methodology with related deliverables, specifically D4.3 and define the following terms.
• **Descriptor** is used to denote a number, semantic label or a vector or matrix of numbers or semantic labels, that characterises a specific aspect of a musical piece. Examples would be the predominant pitch (number) or its evolution (vector of numbers) representing the predominant melody, the key of the piece (semantic label) and its evolution key progression (a list of local key values).

• **Recording** is an audio recording of a specific music piece.

### 2.4 Convention

We use the following writing conventions:

- **bold** for emphasis.
- **italics** for newly introduced terminology.
- **underlined** for cross-references and references to other documents.
3 ONSET AND NOTE LOCATION

Finding the starting point of musically relevant events, i.e. onsets, is a crucial component in various high-level, semantic feature extraction processes, such as structural analysis, beat detection or tempo estimation.

3.1 State of the art

Existing approaches rely on audio signal processing techniques, such as detecting changes in spectral energy [1][2], pitch detection [3], or exploiting phase information [4][2].

Neural networks have been also used to improve beat picking on manually created onset detection functions [5]. Bidirectional recurrent neural networks (RNN) have been used for the task, too [6].

3.2 Proposed approach

Our novel approach, developed as part of this PHENICX deliverable, extends our previous work on onset detection [7][8] and defines a new state of the art.

The approach borrows from image analysis and relies on the fact that onsets, when viewed in the spectrogram, look like edges, which can be nicely seen from Figure 1.

![Figure 1: Spectrogram showing several note onsets.](image)

Hence, we model the onset detection task as an edge detection task, which is a common problem in image processing. In our latest work within the PHENICX project [9], we use convolutional neural networks (CNNs), a variant of feed-forward neural networks. Here, neurons are spatially organized to form feature maps.

We adapt several parameters of the CNN to our type of input data. In particular, we use rectangular filters on the spectrogram. This is motivated by the fact that temporal resolution is more
important than frequency resolution for the task of onset detection. Hence, filters are wide time range and narrow in frequency range. Similarly, we apply maximum pooling.

We train the CNN on spectrogram excerpts centered on the frame of interest, using excerpts of same frame rate and same number of frequency bands (logarithmically scaled), but different window sizes.

The CNN then learns an onset activation function over time, which is smoothed by convolution with a Hanning window. If the local maximum of the function value falls above a fixed threshold, the respective frame is predicted as containing an onset. To avoid over-fitting and thus obtain better generalization properties, we employ a dropout scheme (50% randomly selected input units are ignored in each training example). Doing so, we achieve an F1-score of 89% on a collection of 102 minutes of music with almost 26,000 annotated onsets, containing monophonic as well as polyphonic music. More details can be found in [9].

3.3 Evaluation

Due to the very time-consuming nature of creating ground truth sets for the task of onset detection (manually and precisely defining the onset in possibly highly complex music is an inevitably hard task), existing datasets are sparse. We tested our approach on a joint onset dataset, drawing from different publications (Holzapfel [2], Bello [10]).

The dataset includes a variety of sources, but foremost classical orchestra instruments, both in monophony and polyphony. For instance, violin, cello, clarinet, piano, but also trumpet, saxophone, guitar, and instruments used in non-western music are included.

On this joint dataset, our CNN-based approaches outperform other recent approaches, both in terms of recall and precision (and consequently in F1-score). While other state-of-the-art methods based on RNNs reach about 87% F1-score, our CNN-based approaches exceeds 90% with additional amelioration (dropout, fuzzy training, linear rectifier - cf. [9]).

While not explicitly tested on large-ensemble orchestra music, performance will likely be very similar (to the polyphonic excerpts in the dataset).
4 MULTI-PITCH AND PREDOMINANT MELODIC LINE ESTIMATION

The extraction of pitch information is arguably one of the most important tasks in automatic music description systems. However, previous research and evaluation datasets dealing with pitch estimation focused on relatively limited kinds of musical data. We aim to broaden this scope, and focus it on the context of PHENICX, by studying melody extraction in symphonic western classical music recordings. This material is characterized by a high number of overlapping sources, and by the fact that the melody is played by different instrumental sections, often alternating within an excerpt. We evaluate the performance of twelve state-of-the-art pitch salience functions, multipitch estimation and melody extraction algorithms when determining the sequence of pitches corresponding to the main melody in a varied set of pieces.

4.1 Introduction

Pitch is considered as a subjective psychoacoustical attribute of sound, but is closely related to the physical concept of fundamental frequency \( f_0 \), and both terms are commonly used as synonymous in the Music Information Retrieval (MIR) literature.

Multi-pitch (multiple \( f_0 \)) estimation can be considered as one of the main challenges in the MIR field, as we need to deal with masking, overlapping tones, mixture of harmonic and non-harmonic sources and the fact that the number of sources might be unknown. Current approaches then focus on two simplified tasks.

The first one is the estimation of multiple \( f_0 \) on simple polyphonies (few overlapping notes). The performance obtained by multipitch estimation methods in recent years in the Music Information Retrieval Evaluation eXchange\(^1\) (MIREX) has remained almost constant, reaching 69% note accuracy for relatively simple music material.

The second one is the estimation of the sequence of \( f_0 \) values representing the pitch of the instrument playing the melody from a polyphonic music recording (e.g. singing voice in popular music), a task commonly denoted as melody extraction \([11]\). Most research and evaluation datasets have been focused on vocal pop and jazz music, achieving up to 85% accuracy, while the performance drops to 68% (MIREX05 dataset) when other musical instruments play the melody.

A main challenge for melody estimation algorithms is is to cope with more complex and varied musical material, with melodies played by different instruments, or instrument sections playing in unison, octave relation, or with harmonized melodic lines. Moreover, melody could be played by alternating instruments and sometimes less predominant in terms of energy. All these challenges are found in symphonic music and are addressed in our work. We study the limitations and challenges posed to state-of-the-art pitch estimation methods in the task of estimating the pitches corresponding to the melody in complex polyphonies.

4.2 Evaluation dataset: definition and annotation

The creation of a dataset for melody extraction in symphonic music has been a challenge, partially due to the lack of a specific methodology when there is not a single instrument playing the melody. Inspired by the definition of melody in \([12]\), we collected excerpts in which human

\(^1\)http://www.music-ir.org/mirex
listeners agreed in their “essence”, that is, the sequence of notes that they hum or sing to represent it.

We focused on symphonies and symphonic poems, ballet suites and other musical forms interpreted by symphonic orchestras, mostly from the romantic period and some classical and 20th century pieces. They were sampled to create excerpts with a dominant melody, with lengths ranging from 10 to 32 seconds. In order to verify that the excerpts contained a clear melody and identify the exact sequence of notes, we collected human annotations by recording subjects singing the melody and then performing manual symbolic transcriptions on agreed melodies. The final collection contains 64 audio excerpts with their corresponding annotation of the melody (in MIDI notes), and will be made available for research purposes. The number of excerpts per composer are: Beethoven (13), Brahms (4), Dvorak (4), Grieg (3), Haydn (3), Holst (4), Mussorgski (9), Prokofiev (2), Ravel (3), Rimski-Korsakov (10), Schubert (1), Smetana (2), Strauss (3), Tchaikovski (2), Wagner (1). Regarding the instrumentation, only in one of the excerpts there is a single instrument (oboe) playing the melody. In the rest of the dataset, it is predominantly played by several instruments from an instrument section, or a combination of sections, even alternating within the same excerpts. Figure 2 (left) illustrates the statistics of the predominant instrumental sections playing the melody. In order to account for the length of the notes, we sampled the MIDI files containing the melody (dividing them into frames), and derived a sequence of MIDI values representing the note of the melody at intervals of 1ms. Figure 2 (right) depicts the distribution of pitches in the dataset. Figure 2 (right) also shows a Gaussian model of the distribution, with mean = 74.1 and standard deviation = 12.1. In this dataset, 93.69% of the frames are labelled as voiced (containing a pitch which belongs to the ‘melody’), while 6.31% are unvoiced (none of the pitches are part of the annotated melody).

The files were converted to mono combining left and right channels before executing the extraction, in order to ensure that all algorithms worked with exactly the same material.

### 4.3 State of the art and proposed approach

The evaluated algorithms have been selected considering mainly their relevance in the state of the art, their performance in MIREX (de-facto standard for comparative evaluation of methods for music description), and availability, ideally as open source software, or by having access to their estimations in our dataset.

**Pitch salience functions** give a prominence value to a $f_0$ of interest in each frame of the audio signal, and ideally show only clear peaks at the frequencies corresponding to the present...
The evaluated pitch salience functions are: Salamon [13], Durrieu [14], Marxer [15], Duan [16] and Cancela [17].

**Multipitch and melody estimation algorithms** commonly start by computing some kind of pitch salience function, and then use perceptual principles or additional musical knowledge (timbre, harmonicity, spectral smoothness, etc.) to separate partials and group salience peaks into streams, or even map them to a given pitched source. They may also perform polyphony estimation or voicing detection, applying different approaches (commonly using a threshold).

The evaluated multipitch methods are: Duan et al. [16], Dressler [18], Benetos and Dixon [19]. Dressler’s approach is a more recent implementation of the method in [18], with the main difference that it outputs more pitches.

The evaluated melody extraction methods are: Salamon and Gómez [13], Dressler [20], Fuentes et al. [21] and Durrieu [22].

A summary of the evaluated methods is provided in Table 1. We use the type (SF: salience function, MP: Multiple Pitch estimation, ME: Melody extraction), and the three first letters of the first author’s surname to refer to a specific method (e.g. SF-SAL refers to the salience function by Salamon).

<table>
<thead>
<tr>
<th>Type</th>
<th>(Pre Proc.)+Transform</th>
<th>Salience / Multi./ Estimation</th>
<th>Tracking</th>
<th>Voicing/ Polyph.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancela</td>
<td>SF</td>
<td>CQT</td>
<td>FChT</td>
<td>-</td>
</tr>
<tr>
<td>Durrieu</td>
<td>SF</td>
<td>STFT</td>
<td>NMF on source/filter model</td>
<td>-</td>
</tr>
<tr>
<td>Klapuri</td>
<td>SF</td>
<td>(AF+NT)+STFT</td>
<td>Periodicity analysis</td>
<td>-</td>
</tr>
<tr>
<td>Marxer</td>
<td>SF</td>
<td>STFT</td>
<td>TR</td>
<td>-</td>
</tr>
<tr>
<td>Salamon</td>
<td>SF</td>
<td>(ELF)+STFT+IF</td>
<td>Harmonic summ.</td>
<td>-</td>
</tr>
<tr>
<td>Benetos</td>
<td>MP</td>
<td>CQT</td>
<td>SIPLCA</td>
<td>HMM</td>
</tr>
<tr>
<td>Dressler</td>
<td>MP&amp;ME</td>
<td>MRFFT+IF</td>
<td>Spectral peaks comparison</td>
<td>Streaming rules</td>
</tr>
<tr>
<td>Duan</td>
<td>MP</td>
<td>STFT</td>
<td>ML in frequency</td>
<td>Neighborhood refin.</td>
</tr>
<tr>
<td>Durrieu</td>
<td>ME</td>
<td>STFT</td>
<td>NMF on source/filter model</td>
<td>Viterbi smoothing</td>
</tr>
<tr>
<td>Fuentes</td>
<td>ME</td>
<td>CQT</td>
<td>PLCA on the CQT (EM)</td>
<td>Viterbi smoothing</td>
</tr>
<tr>
<td>Salamon</td>
<td>ME</td>
<td>(ELF)+STFT+IF</td>
<td>Harmonic summ.</td>
<td>Contour-based</td>
</tr>
</tbody>
</table>

Table 1: Overview of evaluated approaches. Square brackets denote that polyphony estimation is not used in the evaluation.

We additionally evaluated our own novel method (further detailed in [23]), based on combining the output of different pitch estimation algorithms. This method is here named as COMB, and another version using the same neighborhood refinement method as MP-DUA is named RCOMB.

### 4.4 Evaluation strategy

Melody extraction algorithms are commonly evaluated by comparing their output against a ground truth (sequence of pitches that the main instrument plays). The evaluation in MIREX focuses on both voicing detection and pitch detection itself. The former is evaluated using metrics from detection theory, such as voicing recall ($R_{vx}$) and voicing false alarm ($FA_{vx}$) rate. The latter is evaluated using accuracy metrics such as raw pitch (RP) and raw chroma accuracy (RC), or overall accuracy (OAC) which represents the proportion of frames that were correctly labeled in terms of both pitch and voicing. Another metric used in the literature is the concordance measure, or weighted raw pitch (WRP) which linearly weights the score of a correctly detected pitch by its distance in cents to the ground truth pitch. Multiple pitch estimation algorithms are

---

2. [http://mtg.upf.edu/technologies/melodia](http://mtg.upf.edu/technologies/melodia)
4. [https://github.com/wslhfnt/separateLeadStereo](https://github.com/wslhfnt/separateLeadStereo)
evaluated using similar metrics, but adapted to the fact that the ground truth contains more than one pitch per frame.

In our case, the ground truth is a sequence of notes corresponding to the notes of the melody, from which we derived a sequence of pitches at intervals of 1ms. This sequence is then resampled to the frame rate of the evaluated algorithm’s output. In this work, salience functions and multipitch algorithms, are evaluated by their ability to output the ground truth pitch of the melody within the first \(N\) estimates. In practice, for the calculation of raw pitch related measures, from a set of \(N\) output pitches per frame, we use only the pitch (in cents) closest to the ground truth, and evaluate the algorithm with this sequence. For the computation of the and raw chroma accuracy, we create a new sequence of pitches by keeping in each frame the pitch (in cents) which is both correct in chroma (chroma match) and closer in cents to the ground truth, or we set a 0 otherwise. For pitch salience functions we extract the 10 most salient peaks with a minimum difference of a quarter tone between them, and order them by salience. The same occurs with the multipitch algorithms, unless they estimate the polyphony (in which case, they output less than 10 pitches). In the case of MP-DRE, the pitches are not ordered by salience, so we just consider \(N = 10\).

### 4.5 Results

The results of the multipitch and salience functions are depicted in Figure 3, showing the mean raw pitch accuracy over all excerpts for a maximum of \(N = 1, 4\) and 10 estimated pitches. The results of the evaluation for one pitch candidate \((N = 1)\) are more detailed in Table 2. We also computed the measures related with the voicing detection for melody extraction methods. The voicing recall rates and false alarm rates \((R_{vx}, FA_{vx})\) are: ME-DRE (88.3%, 74.1%), ME-SAL (60.4%, 39.7%), ME-FUE (79.5%, 66.0%), and ME-DUR (99.8%, 97.7%).

<table>
<thead>
<tr>
<th></th>
<th>RP</th>
<th>WRP</th>
<th>RC</th>
<th>OAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCOMB</td>
<td>62.3</td>
<td>43.5</td>
<td>77.5</td>
<td>58.2</td>
</tr>
<tr>
<td>COMB</td>
<td>63.9</td>
<td>44.6</td>
<td>79.5</td>
<td>59.7</td>
</tr>
<tr>
<td>SF-CAN</td>
<td>45.4</td>
<td>31.1</td>
<td>65.4</td>
<td>42.6</td>
</tr>
<tr>
<td>SF-DUR</td>
<td>61.8</td>
<td>43.2</td>
<td>76.9</td>
<td>57.8</td>
</tr>
<tr>
<td>SF-KLA</td>
<td>34.1</td>
<td>25.5</td>
<td>63.8</td>
<td>32.1</td>
</tr>
<tr>
<td>SF-MAR</td>
<td>42.0</td>
<td>30.7</td>
<td>68.9</td>
<td>39.3</td>
</tr>
<tr>
<td>SF-SAL</td>
<td>34.5</td>
<td>25.3</td>
<td>62.7</td>
<td>32.4</td>
</tr>
<tr>
<td>MP-BEN</td>
<td>20.8</td>
<td>11.1</td>
<td>44.0</td>
<td>19.6</td>
</tr>
<tr>
<td>MP-DRE</td>
<td>14.6</td>
<td>11.0</td>
<td>31.2</td>
<td>13.6</td>
</tr>
<tr>
<td>MP-DUA-ref</td>
<td>19.3</td>
<td>13.1</td>
<td>42.1</td>
<td>19.2</td>
</tr>
<tr>
<td>ME-DRE</td>
<td>49.6</td>
<td>37.6</td>
<td>66.7</td>
<td>46.1</td>
</tr>
<tr>
<td>ME-DUR</td>
<td>66.9</td>
<td>47.1</td>
<td>80.7</td>
<td>62.7</td>
</tr>
<tr>
<td>ME-FUE</td>
<td>22.2</td>
<td>17.1</td>
<td>50.4</td>
<td>19.8</td>
</tr>
<tr>
<td>ME-SAL</td>
<td>28.5</td>
<td>21.4</td>
<td>57.0</td>
<td>23.5</td>
</tr>
</tbody>
</table>

Table 2: Results of all metrics with a single estimate

Figure 3 shows that melody extraction algorithms, which output only one pitch \((N=1)\), reach up to 66.9% accuracy. Salience functions and multipitch algorithms are evaluated as well with \(N = 4\) and 10. As expected, we observe that increasing the number of output pitches increases the accuracy up to 94.2%, which is further increased with the proposed combination of algorithms \((\text{COMB})\). Refining the estimations allows improving the results with a high number of estimates \((\text{RCOMB})\). The combination additionally provides a more consistent estimation throughout the excerpts, making this approach more reliable.
Discussion and future work

The best performing methods with many pitch candidates are salience functions (SF-) partly due to the fact that they output 10 pitches, while multipitch methods generally output less pitches. The best raw pitch accuracy (RP) results with $N = 10$ are obtained by SF-MAR, closely followed by SF-DUR. With $N = 4$, the maximum RP decreases by 6.4%, obtained by SF-DUR, followed by SF-MAR. SF-SAL achieves very good performance with high $N$. MP-DRE, MP-BEN, SF-KLA, SF-DUA and SF-CAN achieve worse results.

The best results for $N = 1$ are obtained with ME-DUR, partially due to the very good performance of its salience function (SF-DUR) which has relaxed constraints in the source filter model (not specific for singing voice). In the case of the overall accuracy (OAC), ME-DUR also benefits from the fact that it seems to consider nearly all frames as voiced, which is appropriate for this dataset. While SF-SAL achieves very good performance, the whole melody extraction method (ME-SAL) does not. This shows that this rule-based approach (which performs best in MIREX for overall accuracy) is tuned to the pitch contour features of vocal music (pop, jazz), and is not able to generalize to the characteristics of our dataset.

Furthermore, we propose a simple method for the combination of the estimations, which decreases the variance of the results and improves the accuracy.

Future work deals with using the knowledge obtained in the evaluation to enhance the proposed combination method or design a new approach. We also foresee using the combination method for other tasks such as note level audio-score alignment. These results will be partially used to create a music visualization system (WP6) which is able to deal with the challenges of symphonic music.
5  STRUCTURAL AND TONAL ANALYSIS

The characterization of the tonal content and the structure of music pieces are among the most relevant problems in automatic music description in general, and addressed in PHENICX. Most of the works in this respect focuses on popular music, and they have been optimised for these repertoires. Although some of the state-of-the-art methods have been also applied to symphonic examples, their evaluation remains mostly unaddressed. The reason for that is the inherent difficulty of having a proper ‘ground truth’ to compare with. For the same fundamental reasons, the problem of tonal description (of some usefulness) of symphonic music has been barely attempted by the MIR community. The first aim of our work in PHENICX is thus to characterise the specificities of the symphonic repertoire with respect to tonal and structural description. By specific we mean the musical characteristics and derived signal’s properties which are unique to the symphonic music. A second stage will use this knowledge for improving state-of-the-art algorithms with symphonic music.

5.1  Introduction

5.1.1  Tonality and structure

Tonality is a relevant musical aspect of Western music, being particularly important for a vast amount of the mainstream symphonic repertoire, which is mostly bounded to the Classical and Romantic periods. Tonality is a main factor contributing to some musical forms. For instance, the main structure of the so-called ‘allegro-sonata form’, which scaffolds the first movement of many symphonies, is usually built from its tonal plan. This scheme establishes that certain relevant sections (e.g. thematic groups) are referred to each other by specific key relations. However, this tonal plan is generally clear only on the simplest (early) symphonies, which are barely performed today. These general keys often serve more as ‘structural references’ (to be implied by the listener) rather than having large segments of music actually written ‘in’ these tonalities. The actual tonality is often in constant evolution, passing through many short-term keys (referred to as ‘tonicisations’). One exponent of this practice is usually found at the development sections of many allegro-sonata forms, in which the composition is intended to challenge or disorient the listener with frequent and unexpected tonal shifts. By estimating these long-term tonal references, we would provide a means for explaining some aspects of the musical form to the users, which could be conveyed through visualisation (WP6).

5.1.2  Orchestration and structure

The symphonic repertoire is particularly characterised by the richness of its sonority. Aside the possibilities offered by the combination of different instrumental families (woodwinds, brass, strings and percussion), the pitch content is also fully exploited in terms of register (from the lowest to the highest pitch of the combined instruments). This expands the possible ways of orchestrating musical passages, a resource exploited by composers and conductors to convey different sensations. For clarifying our terminology, by ‘instrumentation’ we mean the combination of instruments playing a given music passage. By ‘orchestration’, on the other hand, we mean the instrumentation together with the properties of the pitch content being played. Orchestration is thus a convenient means for clarifying the structure of the music. Much symphonic writing exploits instrumentation/orchestration for creating contrast between different sections, as well as for developing dynamic transformations between them (e.g. ‘orchestral crescendo’). The relation between instrumentation/orchestration and musical structure motivates the research on orchestral description. Moreover, the impact of orchestration in listeners has a strong and direct perceptual basis, as much of its effect does not require musical training to be understood.
For instance, any normal-hearing person would distinguish a solo from a tutti section. Many of the sound effects created by composers are rooted on orchestration, and its effect on general audiences has been extensively exploited (e.g., in movie soundtracks). Additionally, research on orchestration has been barely considered by the MIR community, so it constitutes a novel research path.

5.1.3 Case study

The chosen work for our case study, Beethoven’s 3rd symphony ‘Eroica’, is generally agreed as a pivotal composition between the Classical and Romantic periods. This work also constitutes a paradigm of formal complexity, as evidenced by the vast literature analysing the symphony. It also involves a significant usage of the symphonic resources, yet not exploiting the full possibilities of later works. This constitutes a proper compromise for analysing a variety of problems specific for the symphonic repertoire, yet avoiding a too complex task, which seems unrealistic in the scope of PHENICX. The length of this symphony is also a convenient feature for our purposes, as this duration is comparable to that of the mainstream symphonic repertoire.

We considered a stereo mix of the symphony, performed by the Royal Concertgebouw Orchestra (RCO), to generate the computational descriptors. In order to perform score-informed analyses at the required resolution, UPF manually aligned audio and score information at beat level.

5.2 State of the art

5.2.1 Tonal description

The basic technique underlying most of the chord and key estimation methods can be summarised as follows. First, the score or the audio signal is processed to extract the so-called ‘chroma features’. This feature accounts for the pitch-class content of the music over time, which is the pitch content without octave information. This chroma feature is then segmented and averaged, to create a time series of Pitch-Class Profiles (PCPs). The basic estimation technique compares the resulting PCPs with chord or key templates, which can be obtained from perceptual experiments [24], musicological criteria [25], or trained from labelled datasets. The best comparison is taken as the estimate. ‘Tonal tracking’ is referred to as the estimations of chords or keys over time, and most of the SoA algorithms implement tracking as an improvement of the plain estimation method. This is generally addressed by probabilistic inference, for which Hidden Markov Models or Neural Networks, trained from annotated datasets, are usual implementations [26]. There exist specific tracks in MIREX for evaluating chord and key estimation algorithms. All these evaluations rely on the existence of ground truth annotations. To date, however, the availability of reliable annotations of chord and key tracking for the symphonic repertoire is anecdotal.

5.2.2 Structural analysis

There exist a wealth of research on structural segmentation of music, from both audio and score representations. Like most of the content descriptors, they have been optimised for popular music. Music structure is generally addressed through the concepts of ‘similarity’ and ‘novelty’. Similarity is a broad term which refers to some ‘goodness of fit’ measure between descriptors. The usage of Self-Similarity Matrices (SSM), which compute the similarity between each pairwise combination of analysis frames, is common to many approaches. The input to these data structures is most often based on spectral content (e.g., spectrum magnitude, Mel-Frequency Cepstrum Coefficients) or chroma features [27]. These approaches are of some usefulness...
with popular music, based on very approximate ‘repetitions’ of spectral or tonal content. This is the case of the verse-chorus sections in pop and rock music, in which both instrumentation and harmony are often close to identical in different verses or choruses. Any symphonic music of some sophistication deviates significantly from these characteristics. Aside from some standard structures (such as the ‘rondo’, or the literal repetition of the exposition in the allegro-sonata forms), it is generally unlikely to find close timbral or tonal relations between significant sections, in comparable terms as for popular music. Many symphonies make extensive usage of tonal material restatements (e.g. themes or motives), but they often appear transformed in many varied ways. This is particularly complex in symphonic works from the Romantic period onwards. This limits the practical usage of the most common recurrence-finding methods for the kind of repertoire likely to be found in orchestral concerts.

5.2.3 Instrumentation and orchestration

There exist a wealth of research on instrument recognition, closely related with the problems of automatic transcription, source separation and score alignment. However, to our knowledge, no study has addressed specifically the computational description of orchestration from mainstream symphonic repertoire.

5.3 Methodology

In order to understand the specificities of the orchestral repertoire, we have analysed several structure-related descriptors, all of them related with tonal information. For that, we have analysed both the audio signal and a MIDI encoding of the score from the ‘Eroica’. The explored descriptions are: a) key estimation; b) cadence finding; c) instrumentation; and d) orchestration.

The motivation for analysing instrumentation and orchestration is two-fold. Aside their importance for structural analysis, we aim to characterise their impact on the spectral content of the audio signals, from which the musically meaningful tonal descriptors (chords and keys) are extracted.

5.3.1 Structural ground truth

Given the structural complexity of the symphonic repertoire, we addressed a deeper analysis of the ground truth problem, for which we did not follow the standard practices of the MIR community. Our approach combines our own musicological analysis with a comprehensive survey of the vast literature on the ‘Eroica’, through which we compiled the most relevant discrepancies among 8 renowned music analysts. The ground truth annotation of the structural boundaries was based on a voting system: each boundary was first represented by the number of supporting scholars, and a threshold of 3 for a minimum consensus among scholars was applied to select the relevant ones. We also analysed the problem of assigning labels to the segments and/or boundaries, but this information will be considered in future work. The segmentation problem alone proved complex enough to deserve a careful consideration.

5.3.2 Key estimation

We applied the temporal multi-scale key estimation and representation method proposed in a previous work [28], to both a MIDI encoding of the ‘Eroica’ and the audio signal from the RCO’s performance. The method computes a chroma feature time series from each input. The audio signal is processed by a state-of-the-art chroma estimation algorithm, while the MIDI is analysed directly in symbolic domain. The segmentation and tonal estimation stages are agnostic about
the source of the incoming chroma information. This method constitutes a convenient means for evaluation, as it informs separately about the tonal estimation model proper, and about problems of the audio features (e.g. chroma quality). The MIDI encoding was first time-stretched according to the audio performance, in order to have both versions aligned in time. This was done by manual score alignment at the beat level. Fig. 4 depicts a qualitative comparison of a simple key estimation method from audio chroma features (top) and MIDI (bottom), computed from the exposition of the ‘Eroica’. In these plots, the x-axis represents time, the y-axis represents the analysis time-scale (window size for segmenting the music), and each colour represents a different key. The ground truth structural boundaries are annotated as reference (red vertical lines). A qualitative similarity analysis reveals that the key estimation method from audio can be roughly compared to its counterpart from the score. However, the problem of structural segmentation based on key estimation is clearly manifested for music of this tonal complexity. Several of the ground truth structural boundaries are well defined by the tonal estimation, from both score and audio. However, it is also clear that many other short tonicisations are present as well, and some of the segments are not bounded by tonal shifts. The visualization of key information in relation with structure is being subjected to a user study on visualization in the context of WP6-Task 6.1 (visualization of music pieces) and WP7-Task 7.3 (analysis of test results).

![Figure 4: Key estimation from audio (top) and MIDI (bottom), and ground truth segmentation. Exposition of the 1st Movement of the 'Eroica’ symphony](image)

5.3.3 Cadence finding

Cadential processes are among the most common structure-defining resources. We propose a key-independent cadential analysis, based on the description in terms of transposition-invariant set-classes. The general computational framework is described in [29], and information retrieval applications of this kind are described in [30]. We performed a systematic cadential analysis from the first movement of the ‘Eroica’, inspecting a variety of common cadential sets. Our preliminary analysis confirms that some important cadential procedures in symphonies, contributing to large-scale structures, are beyond the description of cadences as plain sequences of chords, and require sophisticated hierarchical interpretation. It is the case of the controversial definition of the second theme group in the first movement of the ‘Eroica’, for which several -technically speaking- perfect cadences towards the second theme’s tonality occur before the (mostly agreed) establishment of the second theme proper. Similar long and complex cadential procedures are featured by much Romantic repertoire which is not symphony-specific, although the extended possibilities of the symphonic resources favors this practice.
5.3.4 Instrumentation

The analysis of the instrumentation points to potential ways of describing structure in terms of instrumental sonority. We analysed the activation of each instrument from the time-aligned MIDI encoding. Fig. 5 depicts this information, together with the ground truth structural boundaries. This reveals that all the structural boundaries correspond to important changes in the instrumentation. On the other hand, it also reveals information of the content of the segments. For instance, a dialogue between instruments is clearly visible in the 6th section, and several tutti sections are easily localised. This kind of information is being subjected to a user study on visualization in the context of WP6-Task 6.1 (visualization of music pieces) and WP7-Task 7.3 (analysis of test results).

5.3.5 Orchestration

The analysis of orchestration points to additional means for describing structure. Several features related with orchestration were explored. The simplest method, not taking into account the specific instruments used, but focusing on chord spans only, consists on a description over time of the number of pitches, the number of pitch-classes, and the number of active instruments. Fig. 6 depicts this information, computed for the exposition of the ‘Eroica’, together with the ground truth structural boundaries (vertical green lines). This representation informs about chord complexity (number of pitch-classes), octavations (ratio between the number of pitches and pitch-classes), and unisons (ratio between number of voices and pitches). As the exploitation of the vertical sonorities constitutes a relevant aspect of the symphonic writing, certain structural information is revealed by this simple description. This information is also relevant for describing some structural dynamic processes, beyond localising boundaries. For instance, Fig. 6 shows several buildup processes leading to tutti sections, a compositional resource referred to as ‘orchestral crescendo’.

Figure 5: Instrumentation and ground truth segmentation.

Figure 6: Orchestration and ground truth segmentation. Number of pitch-classes (red), number of pitches (black) and number of voices (blue). Ground truth segmentation (green).
5.4 Conclusions and future work

We have proposed strategies to the analysis of tonality and structure, designed or adapted for targeting specific aspects unique to (or specially manifested in) symphonic music, with a focus on Beethoven 3rd symphony as a case study.

We plan to exploit instrumentation and orchestration-related information for a better understanding of chroma estimation from symphonies. This will lead to a symphony-specific chroma optimisation, which would improve the chord and key estimation from audio in these repertories. In doing so, many other chroma-based applications (such as score alignment or pattern detection) might be also improved for the symphonic repertoire. Several visualisations related with these descriptors will be proposed and tested within WP6.
6 CONCLUSION

This deliverable presents the selected approaches for music piece analysis in PHENICX. We have focused on symphonic music, with a special emphasis on the Beethoven 3rd symphony (‘Eroica’) as a case study for meaningful musical descriptors. We have evaluated state-of-the-art approaches to estimate the challenges of our particular repertoire, and have adapted and designed novel techniques for locating note events (onset detection), estimating the main melodic lines and describing the structure, orchestration and tonality of the pieces.

All these descriptors will be exploited in the different use cases of the PHENICX prototype, mainly for visualization (WP6).
7 LIST OF AUTHORS

Juan J. Bosch, Agustín Martorell, Emilia Gómez, Music Technology Group, Universitat Pompeu Fabra (UPF), Barcelona

Markus Schedl, Department of Computational Perception, Johannes Kepler University (JKU) Linz
References


