Automatic Melodic and Structural Analysis of Music Material for Enriched Concert Related Experiences

Juan J. Bosch
Universitat Pompeu Fabra,
Music Technology Group,
Roc Boronat 138, Barcelona
juan.bosch@upf.edu

ABSTRACT

This PhD thesis proposal deals with the automatic analysis of musical audio, focusing on the estimation of the predominant melodic lines, which are used as a basis for extracting musical themes, and (along with other features) for structure recognition. The main focus is set on classical western music in large ensemble settings, which poses interesting research challenges to current state-of-the art algorithms. We will study the limitations of current approaches in this genre, and elaborate specific descriptors and methods, combining audio based analysis with further sources of knowledge and modalities. The creation of appropriate datasets will also be a main aspect, in order to properly evaluate the developed approaches. This work will be used to enrich musical concert related experiences, from music consumers to editors.

Categories and Subject Descriptors

H.5.5 Sound and Music Computing (J.5): *Signal analysis, synthesis, and processing*

General Terms

Algorithms, Theory

Keywords

Multipitch estimation; melody extraction; musical voice; structure recognition; music segmentation; musical theme; classical music

1. INTRODUCTION

Music Information Research (MIR) is the field which covers all the research topics involved in the understanding and modeling of music and that use information processing methodologies [16]. Research on this area started in the 1960's, but has seen a major growth from the 2000's, reflected on the increasing number of participants in specialized conferences and evaluation forums in the field such as the Music Information Retrieval Evaluation eXchange (MIREX) [6]. Over the last decade, we have seen how the performance of the algorithms has been increasing, but most of the tasks seem to have reached an upper bound in their performance. Currently, most MIR algorithms are only focused in audio, but the incorporation of further modalities, such as text, score, video, tags, etc. will improve the understanding of musical data [16]. Music signal analysis has borrowed techniques from speech signal processing, but most approaches needed to consider the intrinsic characteristics of musical audio in the techniques and representations, in order to achieve better performance. Further improvements can be obtained by analyzing the characteristics of specific genres, providing tailored solutions. Knowledge coming from other disciplines, such as psychology and musicology is also necessary to approach music in a more integrated manner.

The goal of this PhD thesis is the extraction of predominant melodic lines and the analysis of structure in musical audio. We will deal with the following challenges:

- Multimodal audio analysis, e.g. with score or video
- Combination of a signal processing front-end with knowledge from music perception and musicology
- Consideration of genre-specific constraints

The main focus is set on classical music in large ensembles, which poses interesting research problems to music analysis algorithms. The complexity of three musical signals is visually depicted in the Figure 1, which shows the evolution of their spectrum along time.

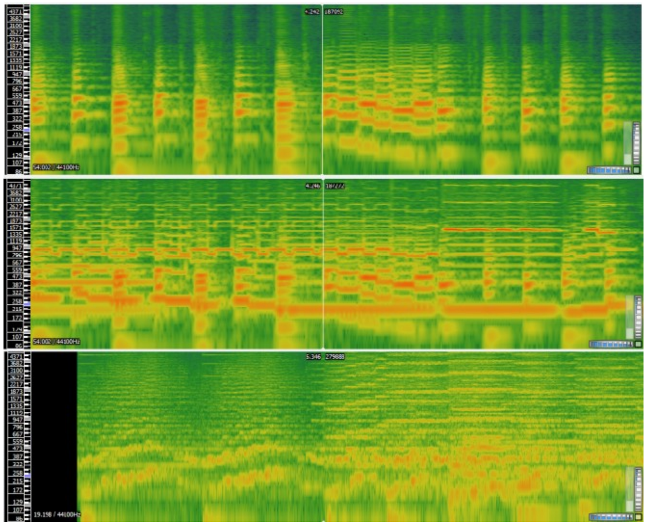

Figure 1: Difference in the spectrogram between a bassoon (top), and bassoon with the rest of a woodwind quintet playing the same piece (center). At the bottom, spectrogram from Smetana's "Ma Blast – Vltava": the first part (bottom-left) shows a section of the orchestra playing in unison. Other sections join (bottom-right) and start playing the main melody

Figure 1 (bottom) corresponds to the case of large ensembles, which in the last seconds presents a high spectral density, due to the high degree of polyphony and the large number of instruments. Additional research problems deal with the subtle

MM '13, Oct 21-25 2013, Barcelona, Spain
ACM 978-1-4503-2404-5/13/10.
http://dx.doi.org/10.1145/2502081.2502218

onsets and offsets, changes in tempo, the lack of a strong beat, complex tonality, and long-term musical structure, with non-obvious boundaries. On the other hand, some benefits of classical music are that the scores are more commonly available, and the musical works may have already fallen in the public domain.

This work is conducted in the scope of the PHENICX project [7], which aims to transform traditional live music concerts and enrich the experiences around them. The research conducted in this PhD thesis will be applied in several demonstrators, visualizing the extracted musical features such as structure, melodic lines, musical themes, etc., adapted to the user's musical knowledge. Concerts can also be experienced at home, where a more detailed analysis of the performance is considered. Additionally, editors which enrich the musical content would benefit of the application of this research, using automatic music analysis for an easier editing, segmentation and music summarization.

2. Scientific Background

This section will introduce the scientific background to be considered in this work. The main aspects are the estimation of multiple pitches, the use of music perception for the analysis of melodic streams and music structure analysis.

2.1 Multiple Pitch Estimation and Musical Stream Analysis

Multipitch estimation deals with the estimation of the fundamental frequencies of several simultaneous sounds. It is the first subtask towards automatic transcription, followed by: onset and offset estimation, instrument identification, and rhythm analysis. An overview of the transcription problem in presented in [10]. A related task is melody extraction, which deals with the estimation of only one pitch, commonly the most predominant. Salamon [14] mentions query-by-humming, karaoke and singer identification as applications of the research in this field, which is mostly focused on human voice.

Many multipitch and melody estimation algorithms start with a time-frequency transformation such as the short time Fourier transform, multi-resolution transforms, or perceptually motivated representations. Some of them use a pre-processing step (as spectral whitening or equal-loudness filters), or frequency refinement as a posterior step. The goal is to extract the (spectral) peaks, and compute the pitch salience, for instance using harmonic summation [9]. Most methods then use perceptual principles [3] or exploit additional musical knowledge such as timbre information, harmonicity (or inharmonicity), spectral smoothness, or synchronous evolution to separate partials, group salience peaks into streams or map them to a given pitched source. Figure 2 shows a salience function and pitch streams for the same orchestral piece as in Figure 1 (bottom), obtained with the MELODIA plugin[1], implementing the work by Salamon [14]. In the case of melody extraction, fundamental frequencies which belong to the melody line are identified by several approaches, e.g. Hidden Markov Models (HMM), dynamic programming, or ad-hoc rules. The last are currently obtaining the best results, especially when using features extracted from the pitch streams [14]. A further additional tracking step is commonly performed in order to correct some extraction errors. While estimating the pitch of a single semi-periodic sound (monopitch estimation) can be achieved with a high accuracy, multipitch estimation is still an unsolved task, especially when several instruments play notes with close harmonic relations.

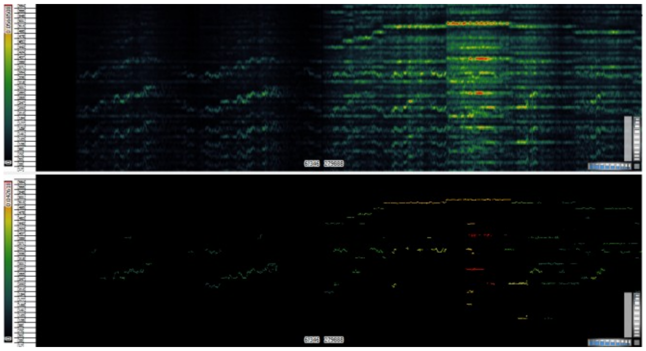

Figure 2: Salience function (top) and pitch streams (bottom) from Smetana's "Ma Blast – Vltava" computed with the MELODIA plugin[1] [14] **for Sonic Visualiser[2].**

The performance obtained in recent years in MIREX has kept almost constant, reaching 69% note accuracy for simple music material. Melody extraction algorithms have however been able to achieve better accuracies (up to 85% in MIREX-2011) [14].

Related work has also been conducted in the symbolic domain, in which several works deal with the separation of a polyphonic music score into musical voices. Commonly, they apply principles derived from auditory scene analysis [3] to group auditory stimuli into streams, as perceived by listeners. Huron [8] presents a review of them, and derives a set of rules to be considered for writing music with perceptibly different voices. Based on this work, Cambouropoulos [4] separated a musical score into voices, using the following principles: "Synchronous Note" (notes starting at the same time are likely to be merged into a voice), "Temporal Continuity" and "Pitch Proximity" (the closer in time and pitch, it is more likely that consecutive notes belong to the same voice). Chew [5] considers slightly different principles, with a different view of musical "voice", not allowing multiple notes in a single stream, as opposed to [4].

2.2 Structure Recognition

Music relies on temporal order, repetition, contrast, variation, and homogeneity of elements such as notes, melodies, chords, harmonies, rhythm, dynamics and timbre. In some periods of western classical music, musical themes have been the material forming the basis of a composition. Commonly, these are melodic elements repeated with some variations.

Structure recognition deals with the division of music into relevant temporal segments, and grouping into relevant clusters. The choice of the features for the automatic analysis restricts the kind of repetitions that can be found. Chroma, timbre, rhythm and dynamics have been mostly used so far, and some approaches use a combination of them. Paulus et al. [13] present a survey of state-of-the-art approaches, and classify them into repetition-based (identifying recurrent patterns), novelty-based (detecting transitions between contrasting parts), and homogeneity-based (finding regions with stationary properties). Serrà et al. [15] combine both homogeneity and repetition. Previous motif extraction approaches in the symbolic domain as [11] work by locating the repeating patterns, selecting those likely to be a theme, and deleting those included in a longer theme. In audio,

[1] http://mtg.upf.edu/technologies/melodia

[2] http://www.sonicvisualiser.org/

the problem is much harder, due to the presence of noise, musical interpretation and expressivity. Weiss [18] considers beat synchronous chromagrams as feature, and a variant of sparse convolutive non-negative matrix factorization. One of the difficulties of structure recognition is that repetitions do not usually take place in precisely the same manner, due to different instrumentation, tempo, dynamics, articulation, as well as melodic and harmonic variations. Muller et al. [12] presented a robust algorithm against such variations, with application to "thumbnailing", a subtask of structure recognition related to music summarization which deals with the determination of segments of audio which best represent the whole piece.

3. APPROACH

In the context of classical music in large ensembles, multipitch analysis becomes intractable with state-of-the-art methods, mainly due to the amount of instruments present and the shared harmonics between pitches. On the other hand, the extraction of a single melody could be too simplistic in certain pieces or styles as counterpoint, where the study of multiple melodic lines becomes necessary. In other orchestral pieces, several instruments contribute to the same melodic stream, which is a case not considered by melody extraction algorithms. The proposed approach considers these facts for the estimation of melodic lines and themes, which are then combined with other musical features for music structure analysis. As presented in the previous section, much research has been conducted in similar directions, but we expect this work to make the following contributions:

- Study the limitations of state-of-the-art multipitch estimation, melody extraction and structure recognition algorithms in orchestral music.
- Elaborate and evaluate appropriate descriptors and methods for genre-specific musical audio analysis, integrating audio based analysis with further sources of knowledge (musicology and psychology), and other modalities, as symbolic (score-based) analysis.
- Create custom and public corpora for the evaluation of the tasks with no appropriate ground truth available.

Figure 3 depicts the main components of the work developed in this PhD thesis, further detailed in the following subsections.

3.1 Multipitch oriented signal processing

Spectral peaks will be first extracted from the audio signal, followed by the creation of a salience function, in a similar manner as [14], but focusing on multipitch estimation instead of predominant melody extraction. We will study the most appropriate pre-processing techniques, time-frequency representations and salience functions.

3.2 Predominant melodic lines estimation

From the pitch salience, we will create pitch contours, following a similar approach to [14], using the principles from Huron [8]. The contours will then be characterized, and grouped into melodic line(s) allowing multiple pitches to contribute to a single auditory stream as in [4]. Three vertical aggregation principles will be considered: "Onset Synchrony", "Tonal Fusion", and "Pitch Co-modulation". The principle of "Tonal Fusion", states that concurrent tones are perceptually less independent if they are separated by intervals (in decreasing order: unisons, octaves, perfect fifths…) that promote tonal fusion.

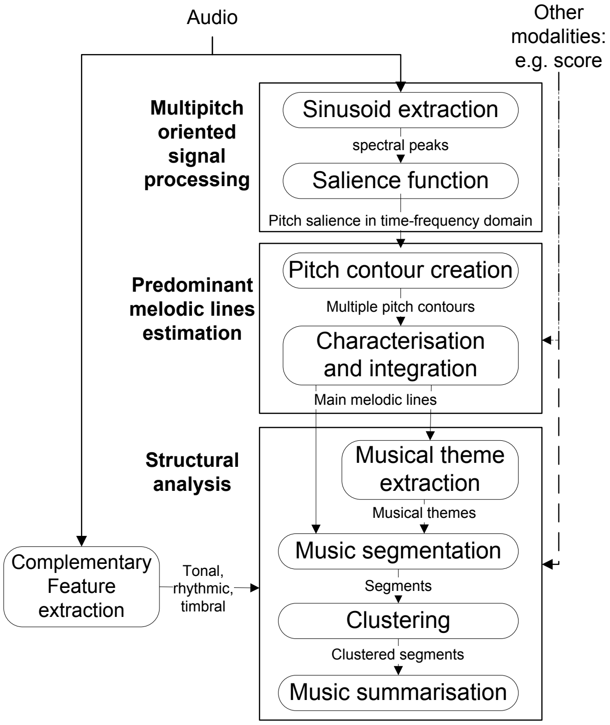

Figure 3: Schema of the main tasks of the PhD

The "Pitch Co-modulation" principle states that concurrent tones are perceptually less independent if their pitch motions are positively correlated. Further knowledge can be exploited, as the timbre, and information from other modalities as the score. Additionally, melodic stream analysis would benefit from the identification of melodic themes and vice versa, since notes within a theme should always belong to the same stream.

3.3 Structural analysis

The combination of timbral, rhythmic, tonal, dynamic, and other descriptors derived from the predominant melodic lines will be investigated for structural analysis. In classical music, an interesting aspect related to the repetition and variations of melodic elements is the identification of musical themes. Appropriate feature combinations and distance measures will be employed for the segmentation, focusing on the robustness against variations in tempo, instrumentation, dynamics, and possible errors in the estimation of melodic elements. The segments will then be clustered and used for creating summaries of the musical piece, considering genre specific characteristics, such as the longer lengths of classical music.

Regarding evaluation, an important corpus for structure recognition is the SALAMI dataset [17], with annotations of classical music from several periods (as well as pop, jazz and other genres). A recently proposed task in MIREX deals with the extraction of musical themes, and could be used for a public evaluation. However, it is currently mostly centered in the analysis of musical works in symbolic domain or from synthesized audio, but not from real recordings, which pose more difficulties for this task, so the creation of a custom dataset will be considered. The definition of musical structure and musical theme has some inherent degree of subjectivity, and the annotation process of the datasets will be carefully examined, or documented in case of a custom corpus.

4. WORK IN PROGRESS

Previous related published work deals with: score-informed melody extraction for predominant instrument source separation [2], and with instrument recognition in polytimbral mixtures [1]. Timber independent and score-informed predominant melody extraction is studied in [2], with application to lead instrument separation based on harmonic and pan-frequency masks. In this work, local and global misalignments between the score and the polyphonic audio mixture are considered. Global misalignments are considered to be due to differences in tempi, including variable tempi. Local misalignments are small misalignments in the onset and offset of the notes, understood as coming from human interpretation, the time envelope of the instrument (attack and decay) or mixing effects. Chroma based Dynamic Time Warping is proposed for the audio-score alignment, with a confidence measure on the score-derived information. The score-based lead instrument pitch tracking starts with a blind estimation of four pitch candidates for the melody. A two-step Viterbi algorithm is then employed to select either one of the candidate frequencies or none, adding probability to the estimated pitch candidates closer to the pitches derived from the score. The proposed solutions improved the estimation of the predominant melody and source separation. In [1] we studied the identification of predominant music instruments, and the benefits of dividing audio into streams, obtaining considerable performance improvement compared to the original instrument recognition algorithm, with several sound segregation algorithms.

Current work deals with the analysis of the problems of state-of-the-art methods for melody and multipitch estimation in large ensemble settings, and the design of a pitch salience function for multipitch estimation. A pitch tracking step will be implemented later, to follow the trajectories of the predominant melodic lines. The evaluation will be conducted against appropriate corpora, including a woodwind quintet, string quartets, and orchestral music recordings.

5. ACKNOWLEDGMENTS

This work is supported by the European Union Seventh Framework Programme FP7 / 2007-2013 through the PHENICX project (grant agreement no. 601166). It was also partially supported by “La Caixa” Fellowship Program.

6. REFERENCES

[1] Bosch, J.J. et al. 2012. A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals. *Proc. ISMIR* (Porto, 2012), 559–564.

[2] Bosch, J.J. et al. 2012. Score-informed and timbre independent lead instrument separation in real-world scenarios. *Proc. EUSIPCO* (Bucharest, 2012), 2417–2421.

[3] Bregman, A.S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press.

[4] Cambouropoulos, E. 2008. Voice And Stream: Perceptual And Computational Modeling Of Voice Separation. *Music Perception*. 26, 1 (Sep. 2008), 75–94.

[5] Chew, E. and Wu, X. 2004. Separating voices in polyphonic music: A contig mapping approach. *Computer Music Modeling and Retrieval*. (2004), 1–20.

[6] Downie, J.S. 2008. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*. 29, 4 (2008), 247–255.

[7] Gómez, E. et al. 2013. PHENICX: Performances as Highly Enriched aNd Interactive Concert Experiences. *Proc. of SMC Sound and Music Computing Conference* (Stockholm, 2013).

[8] Huron, D. 2001. Tone and Voice: A Derivation of the Rules of Voice-Leading from Perceptual Principles. *Music Perception*. 19, 1 (2001), 1–64.

[9] Klapuri, A. 2006. Multiple fundamental frequency estimation by summing harmonic amplitudes. *Proc. ISMIR* (2006), 216–221.

[10] Klapuri, A. et al. 2006. *Signal Processing Methods for Music Transcription*. Springer US.

[11] Meredith, D. et al. 2002. Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*. 31, 4 (2002), 321–345.

[12] Müller, M. et al. 2013. A Robust Fitness Measure for Capturing Repetitions in Music Recordings With Applications to Audio Thumbnailing. *IEEE Transactions on Audio, Speech, and Language Processing*. 21, 3 (Mar. 2013), 531–543.

[13] Paulus, J. et al. 2010. State of the art report: Audio-based music structure analysis. *Proceedings of ISMIR 2010*. (2010), 625–636.

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

[15] Serrà, J. et al. 2012. Unsupervised detection of music boundaries by time series structure features. *Twenty-Sixth AAAI Conference on Artificial Intelligence* (2012).

[16] Serra, X. et al. 2013. *Roadmap for Music Information ReSearch*.

[17] Smith, J.B.L. et al. 2011. Design and creation of a large-scale database of structural annotations. *Proc. ISMIR* (2011).

[18] Weiss, R.J. and Bello, J.P. 2010. Identifying Repeated Patterns in Music Using Sparse Convolutive Non-Negative Matrix Factorization. *Proc. ISMIR* (Utrecht, Netherlands, 2010), 123–128.