FREISCHÜTZ DIGITAL: A MULTIMODAL SCENARIO FOR INFORMED MUSIC PROCESSING

Meinard Müller[1], *Thomas Prätzlich*[1,2], *Benjamin Bohl*[3,4], *Joachim Veit*[3,4]

[1] International Audio Laboratories Erlangen*, [2] Saarland University
[3] University of Paderborn, [4] Hochschule für Musik Detmold

ABSTRACT

In the last decade there has been an explosion in the availability of digitized music material, which comprises data of various formats and modalities including textual, symbolic, acoustic and visual representations. For example, in the case of an opera there typically exist digitized versions of the libretto, different editions of the musical score, as well as a large number of performances given as audio and video recordings. In this paper, we give an overview of various *informed* approaches to music processing, where the availability of multiple sources of music-related information is used for supporting and improving the analysis of music data. Considering the scenario of the opera "Der Freischütz" by Carl Maria von Weber—a work of central musical importance, where one can draw upon a rich body of sources—we highlight how the identification and creation of cross-modal relationships are a key issue in multimedia processing.

Index Terms— Score-informed processing, segmentation, source separation, audio editing, alignment, music synchronization

1. INTRODUCTION

While music research using computers relied in earlier times primarily on symbolic representations of the musical score, the focus of recent research efforts has shifted towards the processing and analysis of various types of music representations including text, audio and video. Actually, music is a challenging multimedia domain offering a multitude of different, complementary representations, and there is a recent trend towards exploiting the abundance of multifaceted information sources for supporting the analysis of music data [1, 2]. Such approaches may also be referred to as *informed* music processing, where information given by one representation (e. g., note parameters specified by a musical score) is used to guide the analysis of another, more complex representation (e. g., the decomposition of an audio recording into different voices). It is such additional information that may pave the way towards solutions of otherwise intractable problems (e. g., blind source separation of polyphonic audio signals).

In this paper, we give an informative overview of recent trends in music processing with a specific focus on the audio domain. Offering a rich body of music-related information and yielding a natural setting for informed approaches, we consider the scenario of the opera "Der Freischütz" by Carl Maria von Weber, see Figure 1. By means of this multimodal scenario, we first discuss a number of challenges that arise in the process of establishing semantic relationships across the various representations. Then, we indicate how the established cross-correlations may open up new ways for music analysis, navigation and retrieval. Doing so, our hope is to indicate

Fig. 1. Music-related information in multiple modalities illustrated by means of the opera "Der Freischütz" by Carl Maria von Weber.

novel, fundamental research directions that draw a wider attraction across different disciplines within and beyond the multimedia community. In the remainder of this paper, we describe the various types of sources that naturally exist in the opera scenario (Section 2) and then discuss how these sources can be used for various music processing tasks related to segmenting, structuring, and decomposing the audio material (Section 3).

2. MUSIC REPRESENTATIONS

As mentioned before, music is an outstanding example since it can be described, represented, and experienced in many different ways [1, 2]. For example, music can be described in *textual* form supplying information on composers, musicians, specific performances, or song lyrics as well as offering detailed descriptions of structural, harmonic, melodic, and rhythmic aspects. Furthermore, *symbolic* representations such as a musical score contains information on the notes such as musical onset time, pitch, duration, and further hints concerning dynamics, agogics, and instrumentation. Finally, *acoustic* representations encode the audio waveform, the actual sound of a specific performance, while *visual* representations show associated information such as live-stage performances or the way musicians physically interact with each other. The complexity of music becomes even more evident when considering that, besides the acoustic domain and the symbolic domain, one also has the auditory or perceived domain that treats music as construct of human minds [3, 4]. Thus, a musical work can be regarded as the entity of all possible ways it can be represented, realized, and perceived.

*The International Audio Laboratories Erlangen (AudioLabs) are a joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS.

In the following, we adopt a more technical point of view and discuss how different music representations naturally appear in various formats and multiple versions in the context of "Der Freischütz".

Opera. Composed by Carl Maria von Weber, "Der Freischütz" is a German romantic opera, which has taken on a key role of musicological and historical importance after its premiere in 1821. Being an opera with interspersed dialogues in the style of the German 'Singspiel', the overture is followed by 16 numbers (arias, duets, trios, instrumental pieces, etc.) [5]. This kind of modular structure allows an opera director for transposing, exchanging and omitting individual numbers, which has led to many different versions and performances.

Text. Besides a rich body of literature on the opera's reception, there is also a detailed account on the libretto by Friedrich Kind [6] and its underlying plot, which is based on an old German folk legend. Since its premiere, the libretto has undergone many changes that were introduced by Kind, not to speak of individual changes made by opera directors. Furthermore, there are versions of the opera in other languages such as French, Russian, or Italian being based on translated versions of the libretto.

Score. Also on the score side, variations have resulted from copying and editing the original autograph score. Changes were not only made by Weber himself, but also by copyists who added further performance instructions and other details to clarify Weber's intention. A scholarly-critical edition of Weber's work[1] keeps track and discusses these variations. The recent Music Encoding Initiative (MEI)[2] aims at developing representations and tools to make such enriched score material digitally accessible. Furthermore, there are various derivatives and arrangements of the opera such as piano transcriptions (e. g., by Liszt) or composed variants of the originally spoken dialogues (e. g., by Berlioz).

Audio and Video. For "Der Freischütz" exist a large number of recorded performances by various orchestras and soloists. For example, the catalogue of the German National Library[3] lists 1200 entries for sound carriers containing at least one musical number of the opera. At least 42 complete recordings have been published and, surely, there still exist many more versions in matters of radio and TV broadcasts. The various performances may reveal substantial differences not only because of the above mentioned variations in the score and libretto, but also because of the artistic freedom taken by opera directors and musicians. Basically everything may vary: numbers are left out, dialogues are shortened or changed, repetitions are omitted, orchestral instrumentation is altered, and so on. Including the visual domain, the differences between opera productions become even more evident in the kind of stage design, not to speak of further derivatives in form of stage musicals and film versions.

3. INFORMED MUSIC AUDIO ANALYSIS

Because of the above mentioned variations with regard to the musical content, the interpretation, and representation as well as the complexity of audio signals of polyphonic music in general, the automated processing of music recordings becomes a challenging area of research. By means of our Freischütz scenario, we now discuss important segmentation and synchronization tasks exploiting the availability of different representations. In Section 3.1, we start with discussing general segmentation tasks that refer to the *temporal* dimension of a given audio data stream. Here, the general goal is to *vertically* partition an audio data stream into musically meaningful sections, each section being specified by a start and end position, see Figure 2. Then, in Section 3.2, we discuss the problem of music synchronization where the objective is to automatically coordinate the multiple information sources of a given piece of music. Finally, in Section 3.3, we address the central problem of audio source separation, where the audio data stream is to be decomposed into meaningful sound events. In the music context, these sound events are typically musical voices, so that source separation can then be regarded as some kind of *horizontal* segmentation, see Figure 3.

3.1. Vertical Segmentation/Structure Analysis

A musical work is typically organized in a hierarchical fashion starting from high-level structures such the movements of a symphony, over mid-level structures such as the parts associated to a musical form, down to low-level structures that correspond to note or even sub-note events. The automatic detection of musically meaningful structural elements constitutes a central task in music processing. Using our Freischütz scenario, we now highlight some high- and mid-level segmentation tasks and principles.

On the high-level side, one first task is to automatically segment a given recorded performance into sections that correspond to the various numbers of the opera. Such sections often correspond to tracks as found on CD recordings of the opera. Using audio material obtained by ripping CD recordings, this problem seems to be already solved. However, in practice one often starts with a single audio file containing the entire opera (e. g., an audio track obtained from a video recording), which then needs to be segmented. Furthermore, the track segmentation found on different CDs is far from being unique. In particular, in some recordings the dialogues appear as separate tracks and in others they are concatenated with musical tracks of the opera's numbers. Finally, different performances of the Freischütz can differ substantially, dialogues are often shortened, and certain numbers may even be missing. In the case that both a reference version and a reference track segmentation of the opera are available, one can use cross-version retrieval techniques to consistently segment other, unsegmented versions. Here, the reference tracks function as "queries" to identify corresponding sections in an unsegmented version, which can then be segmented accordingly. Technically, one requires retrieval methods as known from audio matching [7] or cover song identification [8].

Once the high-level track segmentation has been determined, each track can be further subdivided into mid-level musical sections that reflect the musical form of the respective number. For example, these sections can be the individual stanzas of a given song (as in the example of Figure 2) or the subdivision of the overture into the slow introduction, the fast middle part, and the fortissimo concluding part. In general, the task of recovering a description of the musical form from a given music recording is referred to as *audio structure analysis*, see, e. g., [9, 10]. Following [9], one can distinguish between *repetition-based*, *homogeneity-based*, and *novelty-based* segmentation approaches accounting for different relationships between musical elements based on temporal order, repetition, contrast, variation, and homogeneity. For example, in our opera scenario, homogeneity-based segmentation procedures are required to divide up the audio stream into spoken dialogues and musical sections (related to speech-music classification), to discriminate solo sections from instrumental sections (using timbre features), or to distinguish slow parts from fast parts (using tempo features). Furthermore, repetition-based methods are needed for identifying re-

[1] http://www.weber-gesamtausgabe.de/en/

[2] http://music-encoding.org/

[3] http://www.dnb.de/EN/

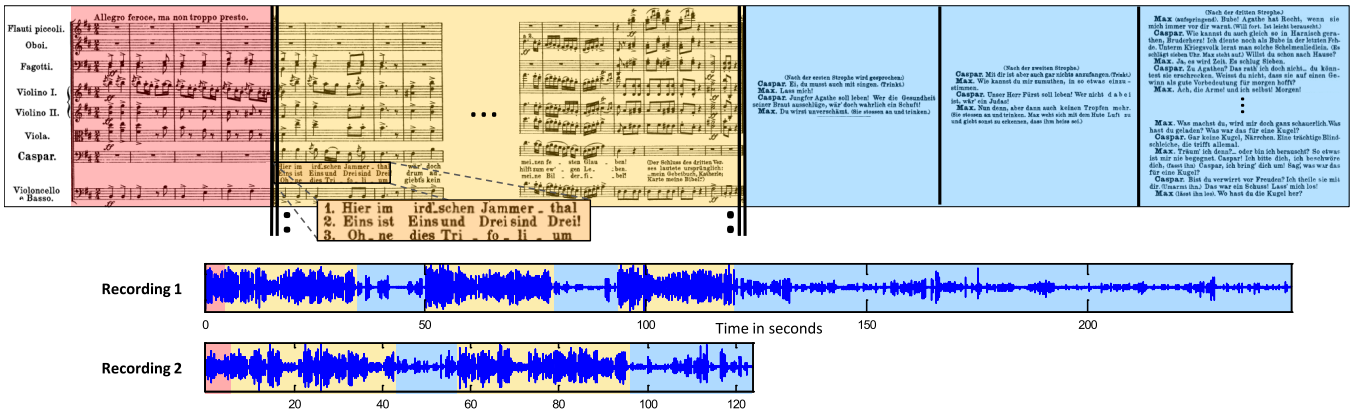

Fig. 2. Different representations of the song "Hier im ird'schen Jammerthal" (No. 4) of "Der Freischütz." **Top:** Score representation. In this song, after an intro ('red'), the repeated stanzas ('yellow') are interleaved with spoken dialogues ('blue'). According to the score, there are three stanzas. **Bottom:** Two different audio recordings of the song. The structure of the second recording deviates from the score by omitting the second dialog and the third verse as well as by drastically shortening the final dialog.

curring patterns that play a crucial role in structuring music. Finally, *novelty-based* methods allow for detecting transitions between contrasting parts such as a piano section followed by a fortissimo section or a change in the musical key (mid-level harmonic change). Figure 2 illustrates a typical mid-level segmentation result. Again, as a major challenge, one needs to deal with musical and acoustic variabilities that may concern the tempo, the sound of instruments, room acoustics, just to name a few. This requires musically informed audio features that capture the respective desired musical properties (e. g. harmony, tempo, timbre), see [9, 10].

3.2. Music Synchronization

In order to exploit the rich body of different sources given for a musical work, one key issue is the development of alignment or synchronization techniques with the aim to identify and link semantically corresponding events present in different representations [11, 12, 13, 14, 15]. Depending on the respective data types, one can distinguish between different synchronization scenarios. For example, in *SheetMusic-Audio* synchronization the objective is to link regions (given as pixel coordinates) within the scanned images of a given sheet music representation to semantically corresponding physical time positions within an audio recording. Such linking structures are useful as navigation and browsing aids in the context of digital music libraries, e. g., to highlight the current position in the scanned score during playback of the recording [13]. In *Audio-Audio* synchronization, the task is to time align two different audio recordings of a piece of music. These alignments can be used to jump freely between different interpretations, thus making efficient and convenient audio browsing possible. Such synchronization tasks are well-defined and tractable as long as the two versions to be aligned are structurally similar so that any event in one version has a counterpart in the other version. Much harder becomes the synchronization if one only has sparse information or partial similarities. For example this is the case in *Lyrics-Audio* synchronization with the goal to align given lyrics to an audio recording of the underlying song, which is useful not only for retrieval but also for karaoke applications. Here, the localization of sung lyrics in complex polyphonic music turns out to be a very hard problem, which can be alleviated when exploiting additional structural or harmonic information [14].

In the context of informed music processing, it is often a score-like symbolic representation that serves as a reference or is used as an additional source of information to facilitate the processing of the audio material. Such strategies are also referred to as *score-informed* approaches. To this end, one needs *Score-Audio* synchronization techniques to establish correspondences between the note events given by the score and time positions in the music recording. As opposed to most navigation and retrieval applications, one needs a much higher degree of temporal accuracy when applying synchronization results in the context of informed music processing. The development of robust high-resolution synchronization techniques constitutes a challenging research direction [12, 15]. Here, one often exploits instrument characteristics (e. g., the attack phase of percussive instruments) to improve the temporal accuracy. In the following section, we discuss a score-informed approach in the context of audio source separation.

3.3. Horizontal Segmentation/Source Separation

The general goal of *source separation* is to decompose an audio signal into its constituent components. In the music context, these components may refer to drum tracks, to the main melody, the bass line, singing voices, or other instrumental voices. One particular challenge arises from the property that the various musical sources are highly correlated with respect to time and frequency [16]. Musicians may follow the same rhythmic patterns and play related melody lines while moving within the same harmonic context. This makes the separation of individual voices from a polyphonic sound mixture an extremely difficult and generally intractable problem. Therefore, when processing music data, musically informed techniques are needed that exploit musical knowledge or music-specific constraints. For example, when extracting the singer's voice in a solo aria, one may exploit that the voice is often characterized by the presence of vibrato (frequency modulations). Or when separating the melody played by a lead instrument, one may exploit its dominance in dynamics and its temporal continuity.

In the last years, so-called *score-informed* approaches have become popular, where the availability of a score representation along with the music recording is assumed [16, 17]. Recent approaches show that the extraction of musical voices from highly overlapping sound sources comes within reach (at least to a certain degree) when exploiting additional musical cues such as timing, pitch, and in-

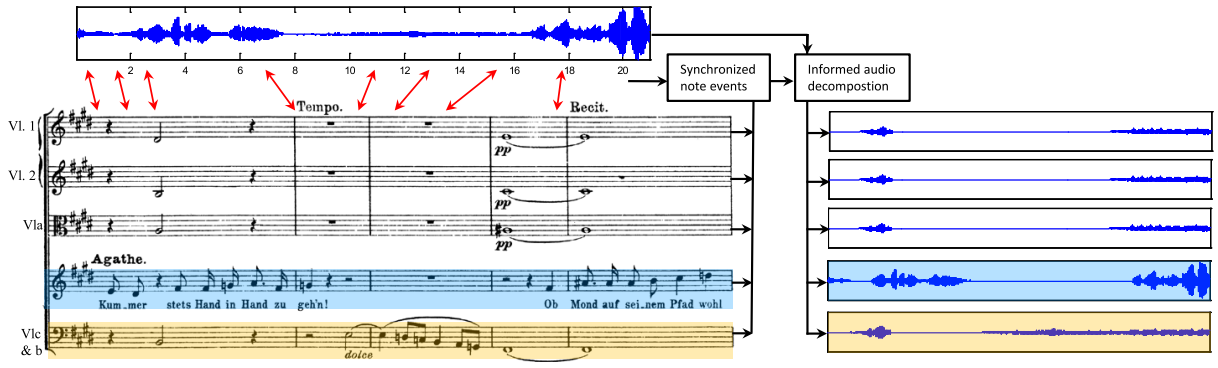

Fig. 3. Illustration of score-informed source separation. After synchronizing the audio recording with the score (the links being indicated by the bidirectional red arrows), the score information is used to guide the decomposition of the audio recording into tracks corresponding to the various voices (e. g., 'blue' indicating the singing voice and 'yellow' the violoncello).

strument information as specified by the score. Additionally, the score representation allows a user to conveniently specify the musical voices to be separated, which can then be used to realize an intelligent equalizer that allows a user to amplify or attenuate certain voices (instead of fixed frequency bands) [18], see also Figure 3 for an illustration.

Even though good progress has been achieved for certain restricted scenarios (e. g. piano music), one still faces numerous challenges in score-informed source separation (not to speak of blind source separation). Obviously, major problems arise when the performer deviates from the score (playing additional notes or leaving out notes). Furthermore, a score does not determine all musical parameters needed to generate a "valid" musical performance and, in practice, a performer has a significant degree of freedom when interpreting a score. Finally, the score does not reflect modulation effects as arising from vibrato or tremolo not to mention acoustic properties introduced by room acoustics (e. g., reverberation) or audio post-processing.

4. CONCLUSION

By means of a concrete opera scenario, we showed how various audio processing tasks arise and how they can be tackled by exploiting complementary information sources associated with the same musical work. However, music processing needs to go far beyond symbolic and acoustic domain by considering also visual representations, the role of the human user, as well as perceptual and subjective aspects. In view of its complexity and richness, music constitutes a challenging test-bed being worth to be considered by the general multimedia community.

5. REFERENCES

[1] Cynthia C. S. Liem, Meinard Müller, Douglas Eck, George Tzanetakis, and Alan Hanjalic, "The need for music information retrieval with user-centered and multimodal strategies," in *Proc. ACM Workshop MIRUM*, 2011, pp. 1–6.

[2] Meinard Müller, Masataka Goto, and Markus Schedl, Eds., *Multimodal Music Processing*, vol. 3 of *Dagstuhl Follow-Ups*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Germany, 2012.

[3] Guerino Mazzola, *The Topos of Music: Geometric Logic of Concepts, Theory, and Performance*, Birkhäuser, 2002.

[4] Geraint A. Wiggins, Daniel Müllensiefen, and Marcus T. Pearce, "On the non-existence of music: Why music theory is a figment of the imagination," *Musicae Scientiae, Discussion Forum*, vol. 5, pp. 231–255, 2010.

[5] John Warrack, *Carl Maria von Weber*, Cambridge University Press, 1976.

[6] Solveig Schreiter, *Friedrich Kind & Carl Maria von Weber - Der Freischütz. Kritische Textbuch-Edition*, Allitera Verlag, 2007.

[7] Frank Kurth and Meinard Müller, "Efficient index-based audio matching," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 382–395, 2008.

[8] Joan Serrà, Emilia Gómez, Perfecto Herrera, and Xavier Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 1138–1151, 2008.

[9] Jouni Paulus, Meinard Müller, and Anssi P. Klapuri, "Audio-based music structure analysis," *Proc. ISMIR*, 2010, pp. 625–636.

[10] Geoffroy Peeters, "Deriving musical structure from signal analysis for music audio summary generation: "sequence" and "state" approach," *Proc. CMMR*, LNCS, Vol. 2771, 2004, pp. 143–166.

[11] Ning Hu, Roger B. Dannenberg, and George Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. IEEE WASPAA*, 2003.

[12] Sebastian Ewert, Meinard Müller, and Peter Grosche, "High resolution audio synchronization using chroma onset features," in *Proc. IEEE ICASSP*, 2009, pp. 1869–1872.

[13] David Damm, Christian Fremerey, Verena Thomas, Michael Clausen, Frank Kurth, and Meinard Müller, "A digital library framework for heterogeneous music collections: from document acquisition to cross-modal interaction," *International Journal on Digital Libraries: Special Issue on Music Digital Libraries*, vol. 12, no. 2-3, pp. 53–71, 2012.

[14] Hiromasa Fujihara and Masataka Goto, "Lyrics-to-audio alignment and its application," *Dagstuhl Follow-Ups*, Vol. 3, pp. 23–36, 2012.

[15] Cyril Joder, Slim Essid, and Gaël Richard, "A conditional random field framework for robust and scalable audio-to-score matching," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2385–2397, 2011.

[16] Sebastian Ewert and Meinard Müller, "Score-informed source separation for music signals," *Dagstuhl Follow-Ups*, Vol. 3, pp. 73–94, 2012.

[17] Romain Hennequin, Bertrand David, and Roland Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. IEEE ICASSP*, 2011, pp. 45–48.

[18] K. Itoyama, M. Goto, K. Komatani, T. Ogata and H. G. Okuno, "Instrument equalizer for query-by-example retrieval: Improving sound source separation based on integrated harmonic and inharmonic models," *Proc. ISMIR*, 2008, pp. 133–138.