Meinard Müller, Thomas Prätzlich, Christian Dittmar

Freischütz Digital
When Computer Science Meets Musicology

1 Introduction

Significant digitization efforts have resulted in large music collections, which comprise music-related documents of various types and formats including text, symbolic data, audio, image, and video. 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 available as audio and video recordings. In the field of music information retrieval (MIR), great efforts are directed towards the development of technologies that allow users to access and explore music in all its different facets. For example, during playback of a CD recording, a digital music player may present the corresponding musical score while highlighting the current playback position within the score. On demand, additional information about the performance, the instrumentation, the melody, or other musical attributes may be automatically presented to the listener. A suitable user interface displays the musical score or the structure of the current piece of music, which allows the user to directly jump to any part within the recording without tedious fast-forwarding and rewinding.

The project *Freischütz Digital* (FreiDi) offered an interdisciplinary platform for musicologists and computer scientists to jointly develop and introduce computer-based methods that enhance human involvement with music. The opera *Der Freischütz* by Carl Maria von Weber served as an example scenario. This work plays a central role in the Western music literature and is of high relevance for musicological studies. Also, this opera was chosen because of its rich body of available sources—including different versions of the musical score, the libretto, and audio recordings. One goal of the project was to explore techniques for establishing a virtual archive of relevant digitized objects, including symbolic representations of the autograph score and other musical sources (encoded in MEI),[1] transcriptions and facsimiles of libretti and other

[1] MEI stands for the *Music Encoding Initiative*, which is an open-source effort to define a system for encoding musical documents in a machine-readable structure. See Andrew Hankinson, Perry Roland

textual sources (encoded in TEI)[2] as well as (multi-channel) audio recordings of the opera. A more abstract goal within the Computational Humanities was to gain a better understanding of how automated methods may support the work of a musicologist beyond the development of tools for mere data digitization, restoration, management, and access.

While computer-aided music research 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.[3] One particular challenge of the project was to investigate how automated methods and computer-based interfaces may help to coordinate the multiple information sources. While our project partners focused on the encoding and processing of text- and score-based representations, our main objective was to research on ways that improve the access to audio-based material. To this end, we applied techniques from signal processing and information retrieval to automatically process the music recordings.

In this paper, having a specific focus on the audio domain, we report on our investigations, results, challenges, and experiences within the FreiDi project from an engineer's perspective. Instead of discussing technical details, our goal is to give an intuitive introduction to the various audio processing tasks that have played an important role in the project. As a second contribution of this paper, we highlight various challenges that arise when (even established) techniques are applied to real-world scenarios. We want to emphasize that it was a great pleasure for us to be part of the FreiDi project. Having partners who were willing to explain their research in simple words, ask questions whenever necessary, carefully listen to each other, while showing mutual respect and interest, we have learned a lot beyond our own research.

At this point, we want to thank Joachim Veit for his invitation to become part of this project. It was his open mindedness and potential to integrate the various perspectives that was one key aspect for making this project a success.

---

and Ichiro Fujinaga, *The music encoding initiative as a document-encoding framework*, in: *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Miami 2011, p. 293–298. See also http://music-encoding.org/ [last accessed: 30 Nov. 2015].

2 TEI stands for the *Text Encoding Initiative* http://www.tei-c.org/ [last accessed: 30 Nov. 2015].

3 Cf. 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: *Proceedings of the International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies (MIRUM)*, 2011, p. 1–6; Meinard Müller, Masataka Goto, and Markus Schedl (eds.), *Multimodal Music Processing*, Dagstuhl 2012 (Dagstuhl Follow-Ups 3).

In the remainder of this paper, we first give an overview of the various types of data sources that played a role in the FreiDi project, where we have a particular focus on the audio material (Section 2). Then, we discuss various audio processing tasks including music segmentation (Section 3), music synchronization (Section 4), voice detection (Section 5), and interference reduction in multitrack recordings (Section 6). For each task, we explain the relation to the FreiDi project, describe the algorithmic approaches applied, discuss their benefits and limitations, and summarize the main experimental results. Finally, in Section 7, we conclude the paper and indicate possible research directions. Parts of this paper are based on the authors' publications,[4] which also contain further details and references to related work.

[4] Cf. Christian Dittmar, Bernhard Lehner, Thomas Prätzlich, Meinard Müller and Gerhard Widmer, *Cross-version singing voice detection in classical opera recordings*, in: *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Malaga 2015, p. 618–624; Christian Dittmar, Thomas Prätzlich and Meinard Müller, *Towards cross-version singing voice detection*, in: *Proceedings of the Jahrestagung für Akustik (DAGA)*, Nuremberg 2015, p. 1503–1506; Meinard Müller, Thomas Prätzlich, Benjamin Bohl and Joachim Veit, *Freischütz Digital: a multimodal scenario for informed music processing*, in: *Proceedings of the International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS)*, Paris 2013, p. 1–4; Thomas Prätzlich, Rachel Bittner, Antoine Liutkus and Meinard Müller, *Kernel additive modeling for interference reduction in multi-channel music recordings*, in: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane 2015; Thomas Prätzlich and Meinard Müller, *Freischütz Digital: a case study for reference-based audio segmentation of operas*, in: *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Curitiba 2013, p. 589–594; Thomas Prätzlich and Meinard Müller, *Frame-level audio segmentation for abridged musical works*, in: *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Taipei 2014, 307–312; Daniel Röwenstrunk, Thomas Prätzlich, Thomas Betzwieser, Meinard Müller, Gerd Szwillus and Joachim Veit, *Das Gesamtkunstwerk Oper aus Datensicht – Aspekte des Umgangs mit einer heterogenen Datenlage im BMBF-Projekt "Freischütz Digital"*, in: *Datenbank-Spektrum*, 15 (2015), p. 65–72.

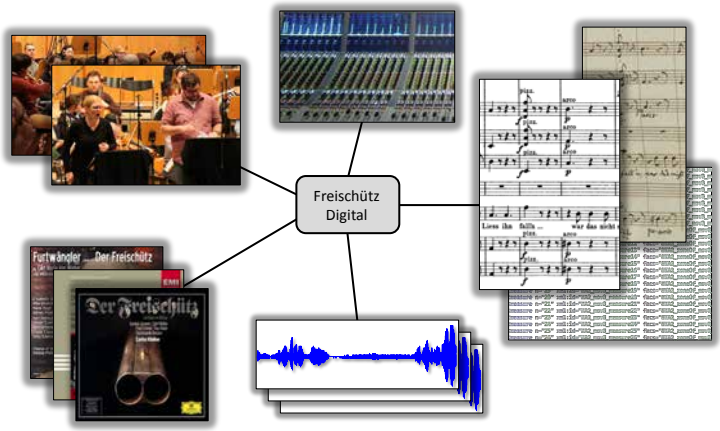


Figure 1: Music-related information in multiple modalities illustrated by means of the opera *Der Freischütz* by Carl Maria von Weber

2 Musical Sources

Music is complex and manifested in many different formats and modalities[5] (see Figure 1). Taking the opera *Der Freischütz* as an example, we encounter a wide variety of multimedia representations, including *textual* representations in form of the libretto (text of the opera), *symbolic* representations (musical score), *acoustic* representations (audio recordings), and *visual* representations (video recordings). In the following, we give some background information on *Der Freischütz* while discussing how different music representations naturally appear in various formats and multiple versions in the context of this opera.

Composed by Carl Maria von Weber, *Der Freischütz* is a German romantic opera (premiere in 1821), which plays a key role in musicological and historical opera studies. The overture is followed by 16 numbers in the form of the German *Singspiel,* where the music is interspersed with spoken dialogues.[6] 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.

---

[5] Cf. Liem et al., *The need for music information retrieval* (see note 3); Müller et al. (eds.), *Multimodal Music Processing* (see note 3).

[6] John Warrack, *Carl Maria von Weber*, London 1976.

As for text-based documents, there are detailed accounts on Friedrich Kind's libretto and its underlying plot, which is based on an old German folk legend.[7] 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. Finally, there exists a rich body of literature on the opera's reception.

On the side of the musical score, there exists a wide range of different sources for the opera. For example, 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[8] keeps track and discusses these variations. The recent *Music Encoding Initiative* (MEI) 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).

As mentioned above, our main focus of this paper is the audio domain. Also for this domain, the opera *Der Freischütz* offers a rich body of available sources including a large number of recorded performances by various orchestras and soloists. For example, the catalogue of the German National Library[9] lists 1200 entries for sound carriers containing at least one musical number of the opera. More than 42 complete recordings have been published and, surely, there still exist many more versions in matters of radio and TV broadcasts. The opera covers a wide range of musical material including arias, duets, trios, and instrumental pieces. Some of the melodic and harmonic material of the numbers is already introduced in the overture. Furthermore, there are numbers containing repetitions of musical parts or verses of songs. The various performances may reveal substantial differences not only because of the above mentioned variations in the score and libretto, but also because a conductor or producer may take the artistic freedom to deviate substantially from what is specified in the musical score. Besides differences in the number of played repetitions, further deviations include omissions of entire numbers as well as significant variations in the spoken dialogues. Apart from such structural deviations, audio recordings of the

---

7 E. g., Solveig Schreiter, *Friedrich Kind & Carl Maria von Weber – Der Freischütz. Kritische Textbuch-Edition*, München 2007.

8 Carl-Maria-von-Weber-Gesamtausgabe, http://www.weber-gesamtausgabe.de/en/ [last accessed: 30 Nov. 2015].

9 http://www.dnb.de/EN/ [last accessed: 30 Nov. 2015].

opera usually differ in their overall length, sound quality, language, and many other aspects. For example, the available recordings show a high variability in their duration, which can be explained by significant tempo differences and also by omissions of material. In particular historic recordings may be of poor acoustic quality due to noise, recording artifacts, or tuning issues (also partly resulting from the digitization process). Working out and understanding the variations and inconsistencies within and across the different sources was a major task we tackled in this project.

3 Track Segmentation

A first audio processing task that emerged in the FreiDi project concerns the automated segmentation of all available audio recordings of the opera in a consistent way. As said, the opera *Der Freischütz* is a number opera starting with an overture followed by 16 numbers, which are interspersed by spoken text (dialogues). When looking at the audio material that originates from CD recordings, the subdivision into CD tracks yields a natural segmentation of the recorded performances. In practice, however, the track segmentations turn out to be rather inconsistent. For example, for 23 different *Freischütz* recordings, Figure 2a shows the track segmentations, which vary between 17 and 41 CD tracks per version. In some recordings, each number of the opera was put into a separate CD track, whereas in others the numbers were divided into music and dialogue tracks, and sometimes the remaining music tracks were even further subdivided. In addition, the CD tracks are often poorly annotated; the metadata may be inconsistent, erroneous, or not available. For digitized material from old sound carriers (such as shellac, LP, or tape recordings), there may not even exist a meaningful segmentation of the audio material. In order to compare semantically corresponding parts in different versions of the opera, a consistent segmentation is needed. In the context of the FreiDi project, such a segmentation was a fundamental requirement for further analysis and processing steps such as the computation of linking structures across different musical sources, including sheet music and audio material (see Section 4).

We presented a reference-based audio segmentation approach,[10] which we now describe in more detail. In our scenario, we assumed that a musicologist may be interested in a specific segmentation of the opera. Therefore, as input of our algorithm, the user may specify a segmentation of the opera by manually annotating the desired segment boundaries within a musical score (or another music representation). This

[10] Prätzlich/Müller, *Freischütz Digital: a case study for reference-based audio segmentation of operas* (see note 4).

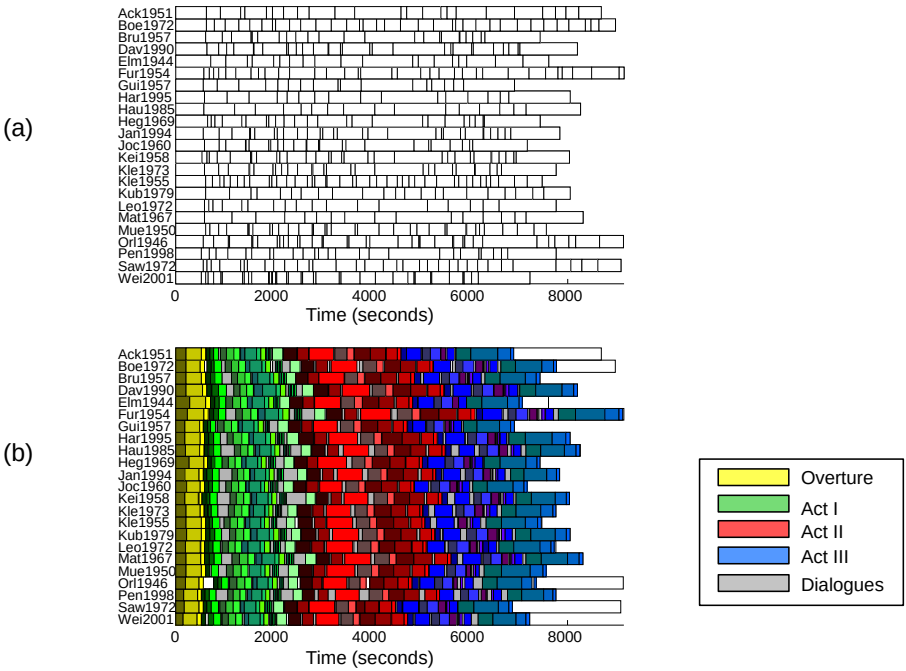


Figure 2: Segmentation of 23 different versions of *Der Freischütz* obtained from commercial CD recordings. **(a)** Segmentation according to the original CD tracks. **(b)** Segmentation according to a reference segmentation specified by a musicologist. The reference segmentation includes 38 musical sections as well as 16 spoken dialogue sections (gray)

annotation is also referred to as *reference segmentation*. For example, in our experiments, a musicologist divided the opera into 38 musical segments and 16 dialogue segments—a segmentation that further refines the overture and the 16 numbers of the opera. Our procedure aims at automatically transferring this reference segmentation onto all available recordings of the opera. The desired result of such a segmentation for 23 *Freischütz* versions is shown in Figure 2b.

As it turned out, the task is more complex as one may think at first glance due to significant acoustic and structural variations across the various recordings. As our main contribution in a case study on recordings of the opera *Der Freischütz*, we applied and adjusted existing synchronization and matching procedures to realize an automated reference-based segmentation procedure.[11] The second and even more important goal of our investigations was to highlight the benefits and limitations of automated procedures within a challenging real-world application scenario. As one main result, we presented an automated procedure that could achieve a segmentation accuracy of nearly 95 % with regard to a suitable evaluation measure. Our approach showed a high degree of robustness to performance variations (tempo, instrumentation, etc.) and poor recording conditions. Among others, we discussed strategies for handling tuning deviations and structural inconsistencies. In particular, short segments proved to be problematic in the presence of structural and acoustic variations.

Another major challenge that turned out in our investigations is the existence of arranged and abridged versions of the opera. In general, large-scale musical works may require a huge number of performing musicians. Therefore, such works have often been arranged for smaller ensembles or reduced for piano. Furthermore, performances of operas may have a duration of up to several hours. Weber's opera *Der Freischütz*, for example, has an average duration of more than two hours. For such large-scale musical works, one often finds abridged versions. These versions usually present the most important material of a musical work in a strongly shortened and structurally modified form. Typically, these structural modifications include omissions of repetitions and other "non-essential" musical passages. Abridged versions were quite common in the early recording days due to duration constraints of the sound carriers. For example, the opera *Der Freischütz* would have filled 18 shellac discs. More recently, abridged versions or excerpts of a musical work can often be found as bonus tracks on CDs.

In our first approach[12] as described above, one main assumption was that a given reference segment either appears more or less in the same form in the unknown

[11] Prätzlich/Müller, *Freischütz Digital: a case study for reference-based audio segmentation of operas* (see note 4).

[12] Ibid.

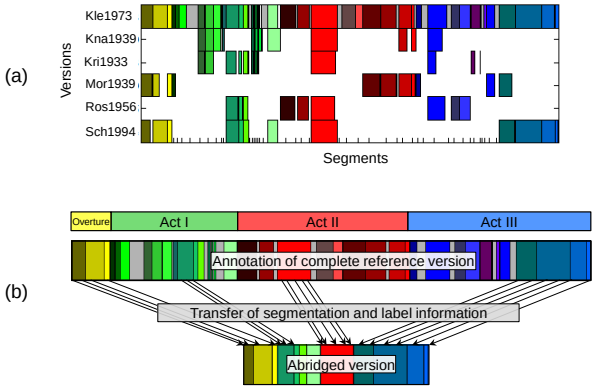


Figure 3: **(a)** Visualization of relative lengths of the segments occuring in abridged versions compared to the reference version "Kle1973". Similar to Figure 2, the gray segments indicate dialogues, whereas the colored segments correspond to musical parts. **(b)** Illustration of the frame-level segmentation pipeline for abridged versions

version or is omitted completely. In abridged versions of an opera, however, this assumption is often invalid. Such versions strongly deviate from the original by omitting material on different scales, ranging from the omission of several musical measures up to entire parts (see Figure 3a). For example, given a segment in a reference version, one may no longer find the start or ending sections of this segment in an unknown version, but only an intermediate section. In a further study, we addressed the problem of transferring a labeled reference segmentation onto an unknown version in the case of abridged versions.[13] Instead of using a segment-based procedure as before,[14] we applied a more flexible frame-level matching procedure. Here, a frame refers to a short audio excerpt on which a suitable audio feature is derived. As illustrated by Figure 3b, the idea is to establish correspondences between frames of a reference version and frames of an unknown version. The labeled segment information of the reference version is then transferred to the unknown version only for frames for which a correspondence has been established. Such a frame-level procedure is more flexible than a segment-level procedure. On the downside, it is less robust. As a main contribution in our study, we showed how to stabilize the robustness of the frame-level matching approach while preserving most of its flexibility.[15]

[13] Prätzlich/Müller, *Frame-level audio segmentation for abridged musical works* (see note 4).
[14] Prätzlich/Müller, *Freischütz Digital: a case study for reference-based audio segmentation of operas* (see note 4).
[15] Prätzlich/Müller, *Frame-level audio segmentation for abridged musical works* (see note 4).

In conclusion, our investigations showed that automated procedures may yield segmentation results with an accuracy of over 90 %, even for versions with strong structural and acoustic variations. Still, for certain applications, segmentation errors in the order of 5 % to 10 % may not be acceptable. Here, we could demonstrate that automated procedures may still prove useful in semiautomatic approaches that also involve some manual intervention.

4 Music Synchronization

A central task in the FreiDi project was to link the different information sources such as a given musical score and the many available audio recordings by developing and adapting synchronization techniques. Generally speaking, the goal of music synchronization is to identify and establish links between semantically corresponding events that occur in different versions and representations.[16] There are many different synchronization scenarios possible depending on the type and nature of the different data sources. For example, in the FreiDi project, there are different versions of the musical score and the libretto (both available as scans and symbolic encodings), as well as a multitude of audio recordings. In *SheetMusic–Audio synchronization*, the task is to link regions of a scanned image (given in pixel coordinates) to semantically

[16] Cf. 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*, in: *International Journal on Digital Libraries: Special Issue on Music Digital Libraries*, 12 (2012), p. 53–71; Sebastian Ewert, Meinard Müller and Peter Grosche, *High resolution audio synchronization using chroma onset features*, in: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei 2009, p. 1869–1872; Hiromasa Fujihara and Masataka Goto, *Lyrics-to-audio alignment and its application*, in: Müller et al. (eds.), *Multimodal Music Processing* (see note 3), p. 23–36; Ning Hu, Roger B. Dannenberg and George Tzanetakis, *Polyphonic audio matching and alignment for music retrieval*, in: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz 2003; Cyril Joder, Slim Essid and Gaël Richard, *A conditional random field framework for robust and scalable audio-to-score matching*, in: *IEEE Transactions on Audio, Speech, and Language Processing*, 19 (2011), p. 2385–2397.

corresponding time positions within an audio recording (specified on a physical time axis given in seconds). In *SymbolicScore–Audio synchronization*, the goal is to link time positions in a symbolic score representation (specified on a musical time axis given in measures) with corresponding time positions of an audio recording (see Figure 4). Similarly, in *Audio–Audio synchronization*, the goal is to time align two different audio recordings of a piece of music.

Two versions of the same piece of music can be rather different. For example, directly comparing a representation of the musical score (that may be given as an XML file) with an audio recording (whose waveform is a sequence of numbers that encode air pressure changes) is hardly possible. In basically all synchronization scenarios, one first needs to transform the given versions into suitable mid-level feature representations that facilitate a direct comparison. The symbolic score, for example, is first transformed into a piano-roll like representation only retaining the notes' start times, durations, and pitches. Subsequently, all occurring pitches are further reduced to the twelve pitch classes (by ignoring octave information). As a result, one obtains a sequence of so-called *pitch class profiles* (often also called *chroma features*), indicating which pitch classes are active at a given point in time. Such features are well suited to characterize the melodic and harmonic progression of music. Similarly, an audio recording can be transformed into a sequence of chroma features by first transforming it into a time-frequency representation. From this representation, a chroma representation can be derived by grouping frequencies that belong to the same pitch class.[17] After transforming both, the score and audio version, into chroma-based representations, the two resulting sequences can be directly compared using standard alignment techniques.[18] In the same fashion, one may also align two audio recordings of the same piece of music (*Audio-Audio synchronization*). Note that this is by far not trivial, since different music recordings may vary significantly with regard to tempo, tuning, dynamics, or instrumentation.

Having established linking structures between musical score and audio versions, one can listen to an audio recording while having the current position in the musical score highlighted.[19] Also, it is possible to use the score as an aid to navigate within an audio version and vice versa. Furthermore, one can use the alignment to seamlessly switch between different recordings, thus facilitating performance comparisons.

---

17 For details see Emilia Gómez, *Tonal Description of Music Audio Signals*. PhD thesis, UPF Barcelona 2006; Meinard Müller, *Information Retrieval for Music and Motion*, Berlin 2007.

18 Müller, *Information Retrieval for Music and Motion* (see note 17).

19 A demonstration of such an interface can be found at http://freischuetz-digital.de/demos/syncPlayer/test/syncPlayer.xhtml [last accessed: 30 Nov. 2015].

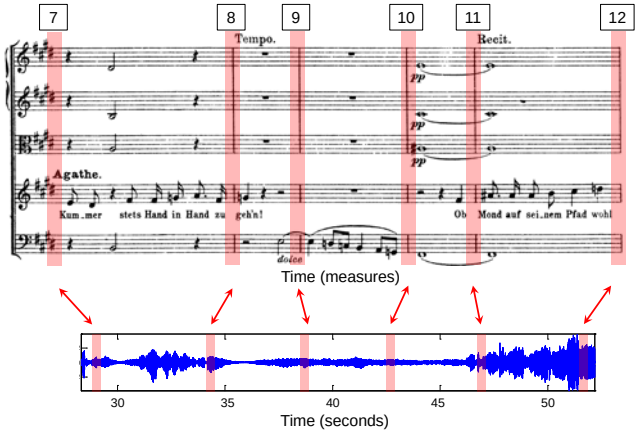


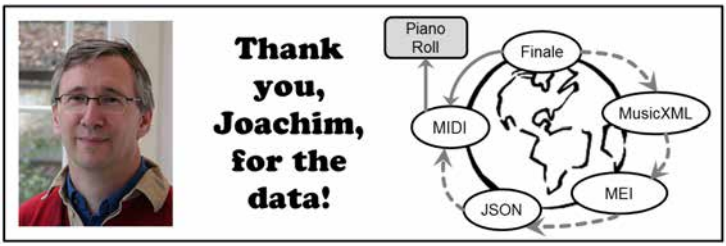
Figure 4: Measure-wise alignment between a sheet music representation and an audio recording. The links are indicated by the bidirectional red arrows

One particular challenge in the FreiDi project are structural variations as discussed in Section 3. In the presence of such variations, the synchronization task may not even be well-defined. Our idea for synchronizing the different versions of *Der Freischütz* was to first use the segmentation techniques from Section 3 in order to identify semantically corresponding parts between the versions to be aligned. This reduces the synchronization problem into smaller subproblems, as only the semantically corresponding parts are synchronized in the subsequent step (instead of the whole opera recordings). Furthermore, since these parts usually have a duration of less then ten minutes, the synchronization procedure becomes computationally feasible even when being computed at a high temporal resolution.

In the case that a reliable prior segmentation is not available, one has to find strategies to compute the alignment even for entire recordings. For example, to synchronize two complete *Freischütz* recordings, one has to deal with roughly five hours of audio material, leading to computational challenges with regard to memory requirements and running time. As one technical contribution within the FreiDi project, we extended an existing multiscale alignment technique that uses an alignment on a coarse resolution to constrain an alignment on a finer grained resolution.[20] In our modified approach,

[20] Meinard Müller, Henning Mattes and Frank Kurth, *An efficient multiscale approach to audio synchronization*, in: *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Victoria 2006, p. 192–197; S. Salvador and P. Chan, *FastDTW: Toward accurate dynamic time warping in*

we proceed in a block-by-block fashion, where an additional block size parameter is introduced to explicitly control the memory requirements. In our experiments, we found that a maximum block size of about eight megabytes is sufficient to yield the same alignment result as a synchronization algorithm without these restrictions. Similar to previously introduced multiscale alignment strategies, our novel procedure drastically reduces the memory requirements and runtimes. In contrast to the previous approach,[21] our block-by-block processing strategy allows for an explicit control over the required memory while being easy to implement. Furthermore, the block-by-block processing allows for a parallel implementation of the procedure.



From a practical perspective, one challenge in the FreiDi project was the handling of the many different formats used to encode symbolic music representations. In view of the alignment task, as mentioned above, we needed to convert the score representation into a piano-roll-like representation which can easily be derived from a MIDI file. In the project, our partners started with an encoding of the score representation using the commercial music notation software *Finale*. The proprietary file format was then exported into MusicXML, which is a more universal format for storing music files and sharing them between different music notation applications. To account for the needs of critical music editions, our project partners further converted the score files into the MEI format which was also chosen to exchange score data within the project. Being a rather new format, only a small number of tools were available for generating, editing, and processing MEI documents. Governed by the limited availability of conversion tools, we exported the MEI files into a JSON representation, which could then be converted into a MIDI representation. Only at the end of the project, we realized that the MIDI export could have been directly obtained by conversion from the original Finale files. From this "detour" we have learned the lesson that there is no format that serves equally well for all purposes. Moreover, the decision for a common file

linear time and space, in: *Proceedings of the KDD Workshop on Mining Temporal and Sequential Data*, 2004, p. 70–80.

21 Müller et al., *An efficient multiscale approach to audio synchronization* (see note 20).

format should be made under careful consideration of the availability and maturity of editing and processing tools.

Even though such experiences are sometimes frustrating, we are convinced that the exploration of novel formats as well as the adaption and development of suitable tools has been one major scientific contribution of the FreiDi project.

5 Dialogue and Singing Voice Detection

As explained in Section 2, the opera *Der Freischütz* consists of musical numbers that are interspersed with dialogues. These spoken dialogues constitute an important part of the opera as they convey the story line. In view of the segmentation and synchronization tasks, knowing the dialogue sections of an opera's recording are important cues. This is illustrated by Figure 5, which shows various representations of the song "Hier im ird'schen Jammerthal" (No. 4). This song consists of an intro (only orchestra) and three verses with different lyrics, but with the same underlying music (notated as repetitions). After each verse, there is a dialogue section. While it is trivial to identify the dialogue sections and the musical structure in a sheet music representation of the song (Figure 5a), this becomes a much harder problem when considering audio recordings of a performance. While the Kleiber recording (Figure 5b) follows the structure as specified in the score, there are omissions in the Ackermann recording (Figure 5c). Knowing the dialogue sections, these structural differences between the two recordings can be understood immediately.

In audio signal processing, the task of discriminating between speech and music signals is a well-studied problem.[22] Most procedures for speech/music discrimination use machine learning techniques that automatically learn a model from example inputs (i. e., audio material labeled as speech and audio material labeled as music) in order to make data-driven predictions or decisions for unknown audio material.[23] The task of speech/music discrimination is an important step for automated speech recognition and general multimedia applications. Within the FreiDi project, we applied and adapted existing speech/music classification approaches to support our segmentation (Section 3) and synchronization approaches (Section 4). Within our opera scenario,

[22] See John Saunders, *Real-time discrimination of broadcast speech/music*, in: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, IEEE, 1996, p. 993–996; Reinhard Sonnleitner, Bernhard Niedermayer, Gerhard Widmer and Jan Schlüter, *A simple and effective spectral feature for speech detection in mixed audio signals*, in: *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, York, UK 2012.

[23] Christopher M. Bishop, *Pattern recognition and machine learning*, New York 2006.

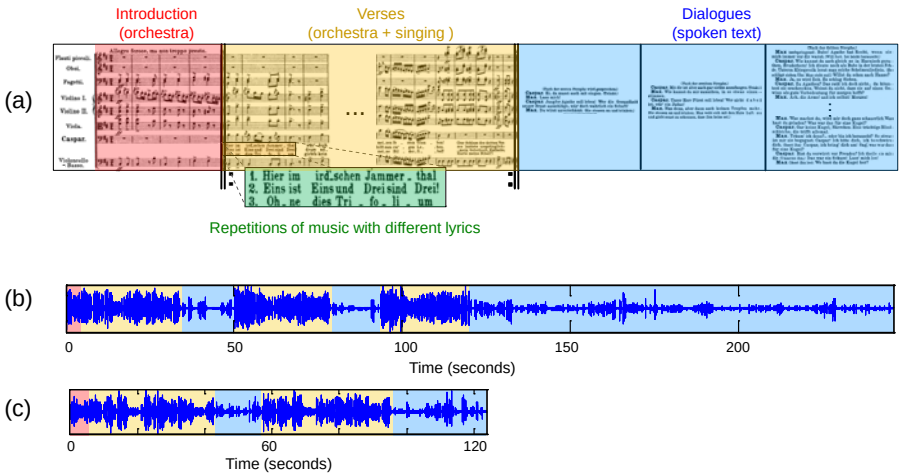


Figure 5: Different representations of the song "Hier im ird'schen Jammerthal" (No. 4) of *Der Freischütz*. **(a)** Score representation. In this song, after an intro (red), the repeated verses (yellow) are interleaved with spoken dialogues (blue). According to the score, there are three verses. **(b)** Waveform of a recorded performance conducted by Carlos Kleiber. The performance follows the structure specified by the above score. **(c)** Waveform of a recorded performance conducted by Otto Ackermann. In this performance, the structure deviates from the score by omitting the second dialogue and the third verse as well as by drastically shortening the final dialogue

it is beneficial to also consider additional classes that correspond to applause and passages of silence. Such extensions have also been discussed extensively in the literature.[24]

A classification task related to speech/music discrimination is referred to *singing voice detection*, where the objective is to automatically segment a given music recording into vocal (where one or more singers are active) and non-vocal (only accompaniment or silence) sections.[25] Due to the huge variety of singing voice characteristics as well as the simultaneous presence of other pitched musical instruments in the accompaniment, singing voice detection is generally considered a much harder problem than speech/music discrimination. For example, the singing voice may reveal complex temporal-spectral patterns, e. g., as a result of vibrato (frequency and amplitude modulations). Also, singing often exhibits a high dynamic range such as soft passages in a lullaby sung in pianissimo or dramatic passages sung by some heroic tenor. Furthermore, many other instruments with similar acoustic characteristics may interfere with the singing voice. This happens especially when the melody lines played by orchestral instruments are similar to the ones of the singing voice.

Technically similar to speech/music discrimination, most approaches for singing voice detection build upon extracting a set of suitable audio features and subsequently applying machine learning in the classification stage.[26] These approaches need extensive training material that reflects the acoustic variance of the classes to be learned. In particular, we used a state-of-the-art singing voice detection system that was originally introduced by Lehner, Widmer and Sonnleitner.[27] This approach employs a classification scheme known as random forests to derive a time-dependent decision function (see Figure 6c). The idea is that the decision function should assume large values close to one for time points with singing voice (vocal class) and small values close to zero otherwise (non-vocal class). In order to binarize the decision function, it

---

[24] Yorgos Patsis and Werner Verhelst, *A speech/music/silence/garbage/classifier for searching and indexing broadcast news material*, in: *International Conference on Database and Expert Systems Application (DEXA)*, Turin 2008, p. 585–589.

[25] Bernhard Lehner, Gerhard Widmer and Reinhard Sonnleitner, *On the reduction of false positives in singing voice detection*, in: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence 2014, p. 7480–7484; Matthias Mauch, Hiromasa Fujihara, Kazuyoshii Yoshii and Masataka Goto, *Timbre and melody features for the recognition of vocal activity and instrumental solos in polyphonic music*, in: *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Miami 2011, p. 233–238; Mathieu Ramona, Gaël Richard and Bertrand David, *Vocal detection in music with support vector machines*, in: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas 2008, p. 1885–1888.

[26] See Lehner et al., *On the reduction of false positives* (see note 25); Mauch et al., *Timbre and melody features* (see note 25); Ramona et al., *Vocal detection* (see note 25).

[27] Lehner et al., *On the reduction of false positives* (see note 25).

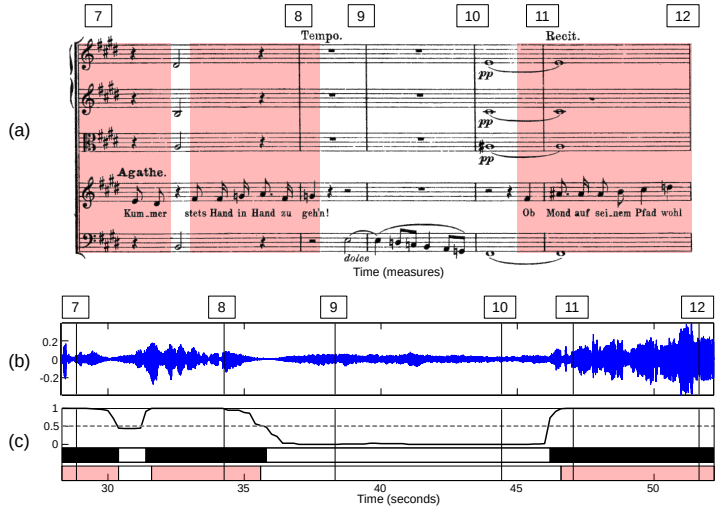


Figure 6: Illustration of the singing voice detection task. **(a)** Score representations of measures 7 to 12 of the song "Wie nahte mir der Schlummer" (No. 8) of *Der Freischütz*. The singing voice sections are highlighted in light red. **(b)** Waveform of a recorded performance. **(c)** Decision function (black curve) of an automated classifier. The function should assume large values (close to one) for time points with singing voice and small values (close to zero) otherwise. The final decision is derived from the curve by using a suitable threshold (dashed horizontal line). The bottom of the figures shows the classification result of the automated procedure (black) and the manually annotated segments (light red)

is compared to a suitable threshold: Only time instances where the decision function exceeds the threshold are classified as vocal.

In particular for popular music, annotated datasets for training and evaluation of singing voice detection algorithms are publicly available.[28] In the context of the FreiDi project, we looked at the singing voice detection problem for the case of classical opera recordings. Not surprising, first experiments showed that a straightforward application of previous approaches (trained on popular music) typically lead to poor classification results when directly applied to classical music.[29] As one contribution, we proposed novel audio features that extend a feature set previously used for popular

[28] Ramona et al., *Vocal detection* (see note 25).

[29] See Lehner et al., *On the reduction of false positives* (see note 25) and our proposed modifications in Dittmar et al., *Cross-version singing voice detection* (see note 4); Dittmar et al., *Towards cross-version singing voice detection* (see note 4).

music recordings. Then, we described a bootstrapping procedure that helps to improve the results in the case that the training data does not match the unknown audio material to be classified. The main idea is to start with a classifier based on some initial training data set to compute a first decision function. Then, the audio frames that correspond to the largest values of this function are used to re-train the classifier. Our experiments showed that this adaptive classifier yields significant improvements for the singing voice detection task. As a final contribution, we showed that a cross-version approach, where one exploits the availability of different recordings of the same piece of music, can help to stabilize the detection results even further.

6 Processing of Multitrack Recordings

In the FreiDi project, a professional recording of No. 6 (duet), No. 8 (aria), No. 9 (trio) of *Der Freischütz* was produced in cooperation with Tonmeister students from the Erich-Thienhaus-Institute (ETI) in Detmold. The main purpose for the recording sessions was to produce royalty free audio material that can be used for demonstration and testing purposes. Furthermore, it was a great opportunity for us to learn about recording techniques and production processes. The generated audio material contains multitrack recordings of the raw microphone signals (one audio track for each microphone) as well as stereo mixes of specific instrument sections and a professionally produced stereo mix of the whole orchestra. Additionally, several variants of the musical score that are relevant for the scholarly edition were recorded to illustrate how these variants sound in an actual performance.[30]

Orchestra recordings typically involve a huge number of musicians and different instruments. Figure 7a shows the orchestra's seating plan, which indicates where each voice (instrument section or singer) was positioned in the room. The seating plan also reflects the number of musicians that were playing in each instrument section. Overall, 44 musicians were involved in the recording session. For large-scale ensembles such as orchestras, interaction between the musicians is very important. For example, each instrument section has a principal musician who leads the other musicians of the section. To make this interaction possible, the different voices are usually recorded in the same room simultaneously. Figure 7b shows the microphones

[30] The recordings are available for download at https://www.audiolabs-erlangen.de/resources/MIR/FreiDi/MultitrackDataset/ [last accessed: 30 Nov. 2015]. For additional audio examples and a further discussion of the production, we refer to http://freischuetz-digital.de/audio-recording-2013.html [last accessed: 30 Nov. 2015] and http://freischuetz-digital.de/audio-production-2014.html [last accessed: 30 Nov. 2015]. See also Johannes Kepper, Solveig Schreiter and Joachim Veit, *Freischütz analog oder digital – Editionsformen im Spannungsfeld von Wissenschaft und Praxis*, in: *editio* 28 (2014), p. 127–150.

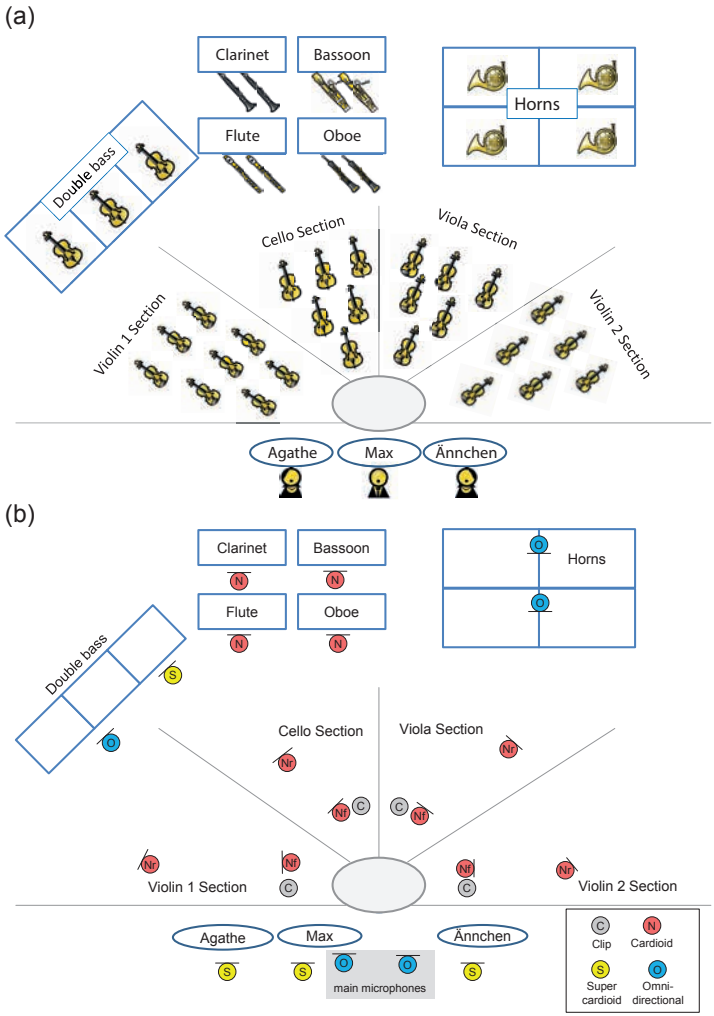


Figure 7: Recording setup used in the FreiDi project. **(a)** Seating plan (German/European style). **(b)** Setup of the 25 microphones used in the recordings, involving two main microphones for recording a stereo image and at least one spot microphone for each instrument section. For each string section, a spot microphone at the front (Nf) and at the rear (Nr) position was used. Additionally, clip microphones (C) were used for principal musicians of the string sections

used for the different voices and their relative position in the room.[31] Two main microphones were used for recording a stereo image of the sound in the room. For capturing the sound of individual voices, at least one additional spot microphone was positioned close to each voice. For some of the instrument sections, additional spot microphones were used, see Figure 7b. The first violin section, for example, was recorded with three microphones: one at the front position, one at the rear position, and a clip microphone attached to the principal musician's instrument. The audio tracks recorded by the spot microphones allow a sound engineer to balance out the volume of the different voices in the mixing process. Usually, a voice is captured by its spot microphones before it arrives at the main microphones which are positioned further away. Therefore, it is important to compensate for different runtimes by delaying the spot microphones such that their signals are synchronized to the main microphones. This avoids unwanted reverberation or artifacts (caused by phase interference) in the mixing process. Furthermore, individual equalizers are applied to each of the microphones to suppress frequencies that are outside of the range of their associated voice.

In such a recording setup, a piece of music is usually recorded in several takes. A take refers to a preliminary recording of a section, that typically covers a few musical measures up to the whole piece. An audio engineer then merges the best combination of takes for the final production. This is done by fading from one take into another at suitable positions in the audio tracks. The merged takes are then used to produce a stereo mixture.[32] In our case, additional stereo mixes that emphasize different aspects of the piece of music were produced. First, a stereo mixture including all voices and microphones was produced. This is the kind of mixture one usually finds in professionally produced CD recordings. For demonstration purposes, additional stereo mixtures were produced for each individual voice (see Figure 7a), as well as for instrument groups including the woodwinds (bassoon, flute, clarinet, oboe), the strings (violin 1, violin 2, viola, cello, double bass), and the singers.

In a typical professional setup, the recording room is equipped with sound absorbing materials and acoustic shields to isolate all the voices as much as possible. However, complete acoustic isolation between the voices is often not possible. In practice and as depicted in Figure 8a, each microphone not only records sound from its dedicated voice, but also from all others in the room. This results in recordings that do not

[31] For No. 6 and No. 8, 23 microphones were used to record 11 voices. For No. 9, 24 microphones were used to record 12 voices.

[32] After merging the different takes, the resulting raw audio material as well as the versions with delay compensation and equalizers were exported for each microphone. In the remaining mixing process, only the versions with delay compensation and equalizers were used.

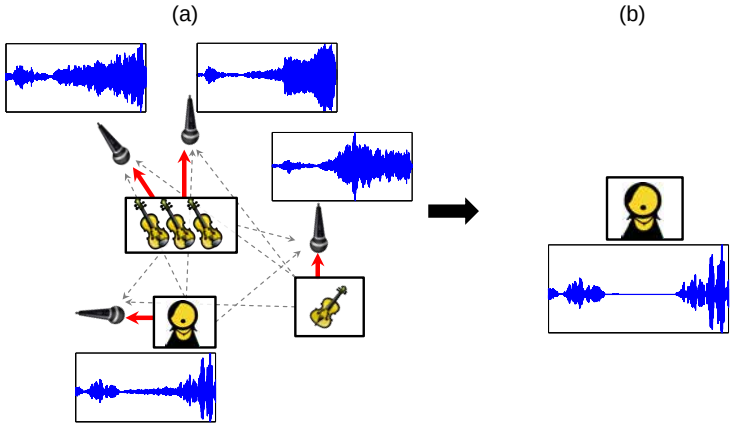


Figure 8: **(a)** Illustration of interference problem in a recording with three voices (violin section, bass, singing voice). A solid line (red) indicates that a voice is associated to a microphone, a dashed line (gray) indicates interference from another voice into a microphone. Each voice is associated with at least one of the microphone channels. **(b)** Interference reduced version of the singing voice signal

feature isolated signals, but rather mixtures of a predominant voice with all others being audible through what is referred to as *interference*, *bleeding*, *crosstalk*, or *leakage*. Such interferences are annoying in practice for several reasons. First, interferences greatly reduce the mixing possibilities for a sound engineer, and second, they prevent the removal or isolation of a voice from the recording, which may be desirable, e. g. for pedagogical reasons or "music minus one" applications (mixtures where a particular voice has been removed). An important question thus arises: is it possible to reduce or remove these interferences to get clean, isolated voice signals? Interference Reduction is closely related to the problem of audio source separation, in which the objective is to separate a sound mixture into its constituent components.[33] Audio source separation in general is a very difficult problem where performance is highly dependent on the signals considered. However, recent studies demonstrate that separation methods can be very effective if prior information about the signals is available.[34]

---

[33] Emmanuel Vincent, Nancy Bertin, Rémi Gribonval and Frédéric Bimbot, *From blind to guided audio source separation: How models and side information can improve the separation of sound*, in: *IEEE Signal Processing Magazine*, 31 (2014), p. 107–115.

[34] See e. g. Antoine Liutkus, Jean-Louis Durrieu, Laurent Daudet and Gaël Richard, *An overview of informed audio source separation*, in: *Proceedings of the International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS)*, Paris 2013, p. 1–4 and references therein).

We recently presented a method that aims to reduce inferences in multitrack recordings to recover only the isolated voices.[35] In our approach, similar to Kokkinis, Reiss, and Mourjopoulos' approach,[36] we exploit the fact that each voice can be assumed to be predominant in its dedicated microphones. Our method iteratively estimates both the time-frequency content of each voice and the corresponding strength in each microphone signal. With this information, we build a filter that strongly reduces the interferences. Figure 8b shows an example of an interference reduced version of the singing voice signal from Figure 8a. Especially in the middle of the corresponding waveforms, it is easy to spot differences. In this region, there was no singing voice in the recording. Hence, the recorded signal in this region originated entirely from interference of other instrumental voices.

In the FreiDi project, we processed the multitrack recordings of the opera to reduce the interferences in the spot microphones.[37] Although the effectiveness of our method has been shown in listening tests, such processings still go along with artifacts that are audible when listening to each interference reduced microphone signal separately. Nevertheless, when using the signals in a mixture, these artifacts are usually not audible as long as not too many voices are drastically lowered or raised in volume. This makes the method applicable in tools like an instrument equalizer where the volume of each voice can be changed separately without affecting the volume of other voices. For example, when studying a specific melody line of the violins and the flutes, an instrument equalizer enables a user to raise the volume for these two voices and to lower it for the others.

7 Conclusions

In this article, we provided a brief overview of our contributions to the FreiDi project, where we investigated how segmentation and synchronization techniques can be used for improving the access to the audio material. For example, automatically computed linking structures may significantly reduce the amount of manual work necessary when processing and comparing different data sources. Furthermore, we showed how automated methods may be useful for systematically revealing and understanding

[35] Prätzlich et al., *Kernel additive modeling for interference reduction in multi-channel music recordings* (see note 4).

[36] Elias K. Kokkinis, Joshua D. Reiss and John Mourjopoulos, *A Wiener filter approach to microphone leakage reduction in close-microphone applications*, in: *IEEE Transactions on Audio, Speech and Language Processing*, 20 (2012), p. 767–779.

[37] Sound examples can be found at http://www.audiolabs-erlangen.de/resources/MIR/2015-ICASSP-KAMIR/ [last accessed: 30 Nov. 2015].

Figure 9: *Freischütz Digital* kick-off meeting in 2012. In the back row: Solveig Schreiter, Raffaele Viglianti, Janette Seuffert, Joachim Veit, Daniel Röwenstrunk, Johannes Kepper; in the front row: Benjamin W. Bohl, Meinard Müller, Thomas Prätzlich (left to right). Missing: Thomas Betzwieser, Gerd Szwillus.

the inconsistencies and variations in the different music recordings. Complementary information sources (such as sheet music and audio recordings) may be exploited to tackle difficult audio processing tasks including singing voice detection and source separation. The multitrack data generated within the FreiDi project can be used as test-bed to study and evaluate such audio processing tasks.

Again, we want to thank Joachim Veit and the entire Freischütz team (see Figure 9) for this very fruitful and exciting collaboration. The FreiDi project has not only indicated how computer-based methods may support musicologists, but also opened up new perspectives of interdisciplinary research between computer scientists and musicologists. With the increase of computing power, the processing of huge audio databases comes within reach. We are convinced that this leads to new ways of computed-assisted research in musicology and the humanities.