FREISCHÜTZ DIGITAL: A CASE STUDY FOR REFERENCE-BASED AUDIO SEGMENTATION OF OPERAS

Thomas Prätzlich
International Audio Laboratories Erlangen
thomas.praetzlich@audiolabs-erlangen.de

Meinard Müller
International Audio Laboratories Erlangen
meinard.mueller@audiolabs-erlangen.de

ABSTRACT

Music information retrieval has started to become more and more important in the humanities by providing tools for computer-assisted processing and analysis of music data. However, when applied to real-world scenarios, even established techniques, which are often developed and tested under lab conditions, reach their limits. In this paper, we illustrate some of these challenges by presenting a study on automated audio segmentation in the context of the interdisciplinary project "Freischütz Digital". One basic task arising in this project is to automatically segment different recordings of the opera "Der Freischütz" according to a reference segmentation specified by a domain expert (musicologist). As it turns 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, we reveal and discuss these variations by systematically adapting segmentation procedures based on synchronization and matching techniques.

1. INTRODUCTION

In recent years, the availability of digital music material has increased drastically including data of various formats and modalities such as textual, symbolic, acoustic and visual representations. 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, which in its totality constitute the body of sources of a musical work. The goal of the ongoing project "Freischütz Digital" [1] is to develop and apply automated methods to support musicologists in editing, analyzing and comparing the various musical sources. The opera "Der Freischütz" by Carl Maria von Weber is a work of central musical importance offering a rich body of sources. Working out and understanding the variations and inconsistencies within and across the different sources constitutes a major challenge tackled in this project. Another more general objective is to apply and to adjust computer-based methods to real-world scenarios and to the needs of domain experts.

One particular problem arising in this case study concerns the automated segmentation of all available audio recordings of the opera. The opera "Der Freischütz" is a *number opera* in the style of a *Singspiel*, starting with an overture followed by 16 numbers (arias, duets, trios, instrumental pieces, etc.) which are interspersed by spoken text (dialogues). In our scenario, the musicologists are interested in a specific segmentation of the opera, which we refer to as the *reference segmentation*. The audio segmentation task is aimed at automatically transferring this reference segmentation onto all available recordings of the opera, see Figure 1 for illustration.

A related scenario is described in [6], where the goal is to identify unknown audio recordings. By applying automated matching procedures, the unknown recordings are compared to well-annotated audio material in a database. Upon identification, the matching result also allows for segmenting the unknown recording. However, this segmentation is more a byproduct, which is not evaluated in detail. In our scenario, the focus lies on the segmentation and, in a certain sense, we follow a reversed approach as we start from known material that we match to a database which we assume to contain representatives of the same musical work.

The contributions of this paper are twofold. First, we apply and adjust existing synchronization and matching procedures to realize an automated reference-based segmentation procedure. The second and even more important goal of this paper is to highlight the various challenges arising in the context of this seemingly easy segmentation scenario. In fact, the various audio recordings reveal significant acoustic and structural deviations. Considering digitized material from old sound carriers (shellac, LP, tape recordings etc.), one often has to deal with artifacts. Structurally, there are omissions or changes of numbers, repetitions, verses and dialogues. By systematically adjusting the segmentation procedure to reveal these variations, we not only successively improve the segmentation quality, but also gain insights into and a better understanding of the audio material.

The remainder of this paper is organized as follows. In Section 2, we describe the various types of sources that naturally exist in the opera scenario and describe the dataset in more detail. In Section 3, we review some basic music synchronization and audio matching procedures. Then, in

[1] http://freischuetz-digital.de/

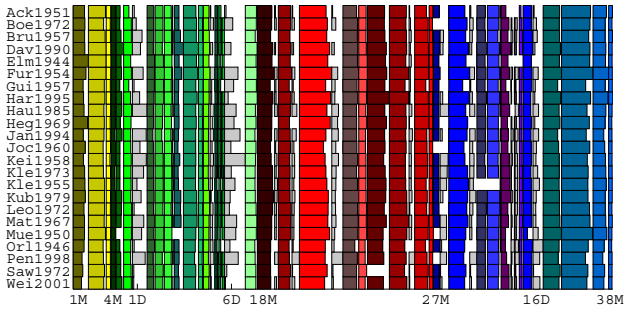


Figure 1. Segmentation result for 23 different audio recordings of "Der Freischütz" according to a reference segmentation specified by musicologists. The reference segmentation includes 38 musical sections (Overture: yellow, Act I: green, Act II: red, Act III: blue) as well as 16 spoken dialogue sections (gray).

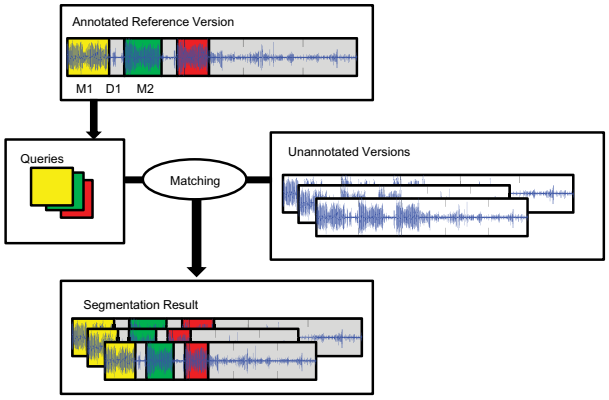


Figure 2. Illustration of the reference-based segmentation procedure.

Section 4, we introduce various segmentation procedures and present a musically informed evaluation of the various results. In Section 5, we conclude the paper and give an outlook to future work. Related work is discussed in the respective sections.

2. MUSICAL BACKGROUND

Music in itself is complex and manifested in many different formats and modalities [5, 9]. For example, for "Der Freischütz" by Carl-Maria von Weber, there are *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, "Der Freischütz" – an important representative of the German romantic opera [11] – serves as a challenging case study. The opera is structured in three acts which are further subdivided into an overture and 16 following numbers interspersed by spoken text passages (dialogues). The numbers cover a wide range of musical material (arias, duets, trios, instrumental pieces, etc.). Some of the melodic and harmonic material of the numbers is already introduced in the overture. Also, some of the numbers contain repetitions of musical parts or verses of songs. In the acoustic domain, these are not always part of the performance, as a the 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 other parts or entire numbers as well as variations in the spoken text and the length of the dialogues. Apart from such structural deviations, audio recordings of the opera usually differ in overall length, sound quality, language and many other aspects. For example, our dataset includes historic recordings that are often prone to noise, artifacts, or tuning problems resulting from the digitization process. Furthermore, the recordings show a high variability in their duration, which can be explained by significant tempo differences and also by omissions of material, see Table 1 and Table 2 for details. Also, there are versions which were adapted into French, Italian and Russian language.

Our raw audio data mostly originates from CD recordings, which were initially segmented in CD tracks, see Table 1. These track segmentations are not consistent, varying between 17 and 41 tracks per recording. In some recordings, each number of the opera was put into a separate track, whereas in others the numbers were divided into music and dialogue tracks, and sometimes the remaining music tracks were even subdivided. In order to compare semantically corresponding parts in different versions of the opera, a consistent segmentation is needed. In the context of the "Freischütz Digital" project, such a segmentation is 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.

In our scenario, a reference segmentation of the musical score into musically meaningful sections was specified by a domain expert (musicologist), who divided the opera into 38 musical segments and 16 dialogue segments. According to this reference segmentation, we manually created an annotation for each of the 23 audio recordings in our database, resulting in over 1000 audio segment, see Figure 1 for an overview. The objective of this paper is to recover this annotation using automated methods and to get a better understanding of the variations and inconsistencies in the audio material.

3. SYNCHRONIZATION AND MATCHING TECHNIQUES

As discussed before, the basic task is to segment an unknown audio recording (assuming no pre-segmentation) according to a given reference segmentation. In the following, we assume that this reference segmentation is specified on the basis of a reference audio recording. Then the objective of the segmentation task is to transfer the segmentation from the reference version to the unknown recording. In this section, we introduce some mathematical notions to model our segmentation problem and then review some standard audio synchronization and matching techniques that are applied in the subsequent section.

Let $X := (x_1, x_2, \ldots, x_N)$ be a suitable feature representation of a given audio recording (the feature type is specified later). Then, a *segment* α is a subset $\alpha = [s:t] \subseteq$

$[1:N] := \{1, 2. \ldots, N\}$ with $s \leq t$. Let $|\alpha| := t - s + 1$ denote the length of α. Furthermore, we define a (partial) *segmentation* of X to be a sequence $(\alpha_1, \ldots, \alpha_I)$ of pairwise disjoint segments, i. e. $\alpha_i \cap \alpha_j = \emptyset$ for $i, j \in [1:I]$, $i \neq j$. Note that in this definition we do not assume that $[1:N]$ is completely covered by the segmentation.

In our scenario we assume that we have a reference sequence X with a reference segmentation $\mathcal{A} = (\alpha_1, \ldots, \alpha_I)$. Furthermore, let $Y := (y_1, y_2, \ldots, y_M)$ be a feature representation of an unknown audio recording. In the case that X and Y are structurally similar on a global scale, the transfer of the reference segmentation of X onto Y can be done by using standard synchronization or alignment techniques [1,3,7]. Here, *music synchronization* denotes a procedure which, for a given position in one representation of a piece of music, determines the corresponding position within another representation. When synchronizing two audio recordings, the first step consists in transforming the recordings into feature representations, typically chroma-based audio features. [2] Based on these feature representations and a suitable cost measure, one applies dynamic time warping (DTW) to compute a cost minimizing warping path which realizes the linking between X and Y, see [7, Chapter 4].

This synchronization-based transfer works as long as X and Y globally coincide. However, problems arise in the presence of significant structural differences. Furthermore, in case X and Y are long (as is the case for complete recordings of entire operas), running time and memory issues arise when performing DTW. Even though (multiscale, forward estimation) acceleration techniques exist [1,10], such techniques are not suited when structural differences occur. As an alternative, one may apply more locally oriented *audio matching* techniques, where the individual segments α_i of the reference segmentation (used as "queries") are matched to subsegments of the unknown sequence Y (resulting in "matches" or "hits"), see [4]. In other words, the cost-intensive global DTW alignment is replaced by several smaller local alignments (realized by a subsequence variant of DTW), see also Figure 2 for illustration. Another positive effect is that using local matches allows for a better handling of missing segments and structural differences. On the downside, by querying the reference segments individually, one may loose temporal coherence, while the chance of obtaining local mismatches is increased (in particular for short segments).

In the subsequent section, we systematically apply, modify and combine both techniques – global synchronization and local matching – for performing our segmentation task. Here, besides the actual segmentation, our main goal is to obtain a better understanding of various kinds of variations and inconsistencies in the audio material.

4. AUDIO SEGMENTATION

In this section, after introducing our evaluation measure to assess the accuracy of segmentation results (Section 4.1), we discuss various strategies to tackle the segmentation task based on global synchronization (Section 4.2) and local matching procedures (Section 4.3 – 4.6). Furthermore, we discuss the benefits and limitations of the respective procedures while revealing the musical and acoustic variations and inconsistencies in the audio material.

4.1 Evaluation Measure

First of all, we need a measure that allows us to compare two given segments α and β. To this end, we define the *relative overlap measure* between α and β to be the value

$$\mu(\alpha, \beta) := \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|} \in [0, 1],$$

which indicates the ratio of the absolute overlap and the length of the union segment. Note that $\mu(\alpha, \beta) = 1$ if and only if $\alpha = \beta$, and $\mu(\alpha, \beta) = 0$ if $\alpha \cap \beta = \emptyset$.

As before, let us assume that the reference version is represented by the sequence $X := (x_1, x_2, \ldots, x_N)$ and the *reference segmentation* by $\mathcal{A} := (\alpha_1, \ldots, \alpha_I)$. Furthermore, let $Y := (y_1, y_2, \ldots, y_M)$ be the unknown version to be segmented. For the purpose of evaluation, we assume that there is also a *ground truth segmentation* $\mathcal{B} := (\beta_1, \ldots, \beta_I)$ for Y, where each β_i musically corresponds to the α_i. The goal is to automatically derive the segmentation of Y. Let P denote such a segmentation procedure, which automatically transfers each reference segment α_i to a computed segment $\mathrm{P}(\alpha_i) \subseteq [1:M]$. Then, the relative overlap measure $\mu(\beta_i, \mathrm{P}(\alpha_i))$ indicates the segmentation quality of the procedure P.

Because of the mentioned structural variations, the version Y does not necessarily contain a segment that musically corresponds to a reference segment α_i. In this case, the ground truth segment is set to $\beta_i = \emptyset$. Furthermore, the procedure P does not have to output a computed segment, which is modeled by setting $\mathrm{P}(\alpha_i) = \emptyset$. In the case that both the segment $\mathrm{P}(\alpha_i)$ and β_i are empty, we define $\mu(\beta_i, \mathrm{P}(\alpha_i)) = 1$ (a non-existing segment has been identified as such). Note that if only one of the segments is empty, $\mu(\beta_i, \mathrm{P}(\alpha_i)) = 0$.

4.2 Global Approach ($\mathrm{S1}, \mathrm{S2}$)

In the following matching procedures and evaluation, we only consider the musical sections (indicated by the non-gray segments in Figure 1) while leaving the dialogue sections (the gray segments in Figure 1) unconsidered. Exemplarily, we use a reference segmentation $\mathcal{A} = (\alpha_1, \alpha_2, \ldots, \alpha_{38})$ based on the recording conducted by Carlos Kleiber in 1973 (`Kle1973`), which is a performance that closely follows the musical score. Quantitative results for all procedures to be discussed are presented in Table 1 (relative overlap averaged over versions) and Table 2 (relative overlap averaged over segments).

In the two procedures $\mathrm{S1}$ and $\mathrm{S2}$, we apply a global synchronization approach. For $\mathrm{S1}$, we employ DTW using the step size condition $\Sigma_1 = \{(1,1), (1,2), (2,1)\}$, see [7, Chapter 4]. This strategy is usually very robust as long as there are no significant deviations in structure

[2] In our experiments, we use chroma-based CENS features of 2 Hz resolution as supplied by the chroma toolbox [8].

and tempo between the two versions compared. However, the procedure S1 is not able to compensate well for structural variations leading to an average relative overlap of 0.852, see Table 1 When using the step size condition $\Sigma_2 = \{(1,1),(1,0),(0,1)\}$ (calling this procedure S2), performance improves significantly, yielding the average relative overlap of 0.930, see Table 1. For example, in the version `Saw1972`, the dialogues are comparatively short, see also the gray rectangles in Figure 1. Such a situation causes S1 to fail, resulting in an overlap of 0.615 compared to 0.896 for S2, see Table 1. For both procedures, the alignment accuracy for α_{38} is very low with 0.714 (S1) and 0.724 (S2), see Table 2. This is due to audio material not belonging to the actual opera that is appended at the end (CD bonus tracks) in some versions. In this case, the global synchronization procedures do not allow to skip the final tracks. Despite the promising results of S2, this approach has several limitations. First, it is inefficient considering runtime and memory requirements, especially when increasing the feature resolution, see also Section 3. Secondly, it is not well suited to accommodate for structural changes in a controlled manner. And thirdly, the procedure does not give deeper insights into the musical and acoustic properties of the underlying audio material.

Our goal in the following sections is to develop a more flexible segmentation strategy that achieves a quality comparable to S2 while yielding better insight into the versions' properties.

4.3 Local Approach (M1)

The remaining approaches discussed below rely on a local matching procedure based on a subsequence variant of DTW using the step size condition Σ_1. Here, for each $\alpha_i \in \mathcal{A}$ (used as a query) applied to a given unknown version, we compute a ranked list of *matching candidates*. For the segmentation procedure M1, we only consider the top match in the list, see also Figure 2 for illustration of the general matching strategy.

In Figure 3a, the relative overlap values for M1 computed on all recordings in our dataset are presented in a gray-scale matrix visualization, where the rows indicate the audio versions and the columns indicate the segments. Black corresponds to $\mu = 0$ (no overlap) and white to $\mu = 1$ (perfect overlap). Row-wise, the segmentation accuracy of a specific version becomes obvious, whereas column-wise, segments which are problematic across versions can easily be spotted. An example for a problematic version is `Elm1944`, which generally seems to perform poorly, showing many black entries in Figure 3a and having a low average relative overlap of 0.705, see Table 1. A closer look at the audio material revealed that there are some issues concerning the tuning of this version, probably resulting from the digitization process. Furthermore, there are segments which show a poor segmentation accuracy across versions, see for example the black entries for α_{14} to α_{16} in Figure 3a. It turns out that these three segments correspond to the three verses of a song (No. 4) in the opera. The reason why this song has been divided into individual segments is that there are dialogues between the verses (recall that a requirement of the reference segmentation was to separate music and dialogue sections). The verses all share the same melodic and harmonic material and are thus easily confused with each other in the matching procedure. Another interesting problem appears for α_{32}, where M1 nearly fails for every version, resulting in an overall segmentation accuracy of 0.157, see Table 2 and Figure 3a. Actually, α_{32} (having a duration of only 12.4 seconds) is a short snippet of a chorus section for which many repetitions exist in the surrounding segments α_{31} (song with several verses and chorus sections) and α_{33} (chorus) which are interspersed by dialogues. Thus it is very likely that α_{32} is matched into the harmonically similar parts within α_{31} or α_{33}. For the version `Kle1955`, segment α_{38} seems to be problematic, see Figure 3a. Actually, α_{38} contains musical material which is already used in the overture of the opera (covered by α_3). A closer look into the matching results for `Kle1955` revealed that α_{38} matched indeed into the musically very similar section in the overture.

In conclusion, procedure M1 is more efficient[3], see also Section 3, while its main drawback is the loss of robustness due to confusion of local matches.

4.4 Tuning Issues (M2)

In real world scenarios, the tuning of a music performance often slightly deviates from the standard tuning, where a chamber tone of 440 Hz serves as reference frequency. This usually influences pitch related audio features such as chroma features. To compensate for different tunings, one typically integrates a tuning estimation procedure in the feature extraction process [2]. In the previous approaches, we already used tuned chroma features. But since an unkown version of the opera also contains a lot of non-music material (dialogues, applause, etc.), which is also considered in the tuning estimation, the resulting estimate may be incorrect.

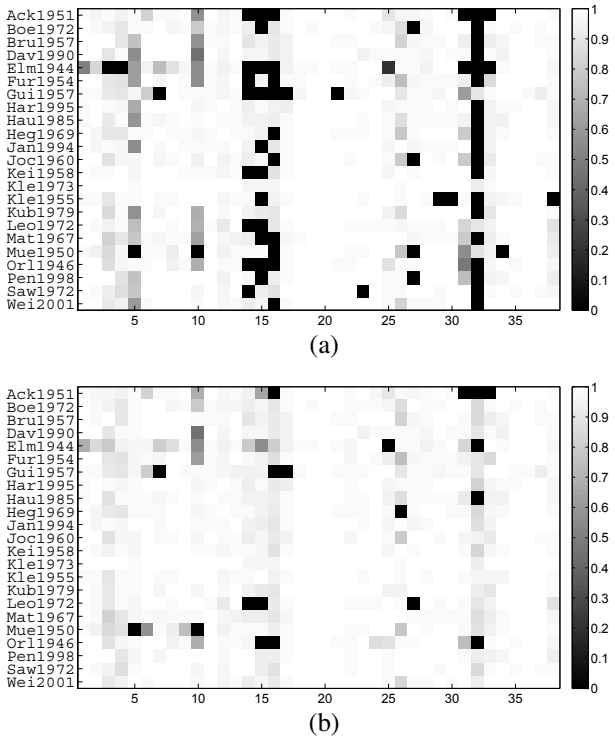
With procedure M2, we evaluate the influence of the tuning estimation on the matching procedure. This problem can either be addressed on the side of the unknown version or on the query side. In our approach, we use the same chroma sequence for the unknown version as in M1, and simulate the tuning deviations on the query side by computing the chroma sequence for the query with respect to six different reference frequencies (in the range of a semitone). Doing this for each query α_i, we then use the chroma sequence yielding the minimum cost in the matching.

For `Elm1944`, the local tuning adjustment indeed leads to a substantial improvement from 0.705 (M1) to 0.777 (M2), see Table 1. Also, there are improvements for certain segments, e.g., α_{38} with 0.921 (M1) compared to 0.968 (M2), see Table 2. In this example, the improvement

[3] On a 64bit machine, the average memory requirement for a global DTW run on one piece of our dataset is 1.7 GB (2 Hz feature resolution) and 42.6 GB (10 Hz), computed from the length of the reference version and the average version length. Upper bounds for the local matching approaches (derived from the maximum query length and the average version length) are 114 MB (2 Hz) and 2.9 GB (10 Hz).

Version	#O	$dur.$	S1	S2	M1	M2	M3	M4
Ack1951	19	6904.81	0.596	0.811	0.808	0.851	0.850	0.853
Boe1972	30	7771.77	0.784	0.931	0.889	0.865	0.962	0.962
Bru1957	24	7439.33	0.906	0.933	0.927	0.905	0.923	0.966
Dav1990	30	8197.88	0.972	0.984	0.926	0.926	0.950	0.961
Elm1944	19	7081.52	0.698	0.827	**0.705**	**0.777**	0.806	0.865
Fur1954	34	9121.69	0.923	0.936	0.866	0.861	0.938	0.949
Gui1957	18	6911.30	0.908	0.937	0.801	0.851	0.860	0.886
Har1995	17	8044.99	0.974	0.981	0.944	0.943	0.965	0.973
Hau1985	17	8245.23	0.955	0.957	0.935	0.933	0.932	0.943
Heg1969	25	7436.75	0.896	0.958	0.913	0.895	0.943	0.946
Jan1994	30	7843.21	0.881	0.987	0.916	0.917	0.964	0.976
Joc1960	26	7178.21	0.922	0.948	0.887	0.911	0.968	0.967
Kei1958	32	8043.00	0.886	0.965	0.904	0.902	0.976	0.975
Kle1973	29	7763.00	1.000	0.996	0.989	0.990	0.990	0.990
Kle1955	41	7459.35	0.776	0.873	0.849	0.876	0.980	0.980
Kub1979	23	8044.65	0.959	0.985	0.927	0.929	0.953	0.974
Leo1972	19	7726.17	0.861	0.926	0.905	0.900	0.875	0.896
Mat1967	17	8309.35	0.984	0.983	0.874	0.876	0.948	0.965
Mue1950	35	7559.97	0.814	0.881	0.825	0.824	0.885	0.895
Orl1946	32	7368.58	0.559	0.807	0.853	0.854	0.852	0.883
Pen1998	26	7768.00	0.866	0.904	0.890	0.891	0.968	0.977
Saw1972	29	6871.02	**0.615**	**0.896**	0.893	0.894	0.968	0.974
Wei2001	38	7220.13	0.859	0.974	0.915	0.916	0.965	0.975
∅	26	7665.65	**0.852**	**0.930**	0.884	0.891	0.931	0.945

Table 1. Relative overlap values averaged over segments for different versions and different procedures. The first column indicates the version, the second (#O) the number of segments on the original sound carrier, and the third column ($dur.$) the overall duration in seconds of the recording. S1, S2, M1, M2 M3, and M4 denote the respective segmentation procedures.



(a)

(b)

Figure 3. Matrix visualization of relative overlap values, where the versions correspond to rows and the segments to columns. **(a)**: P = M1. **(b)**: P = M4.

mainly comes from the version Kle1955, where α_{38} is now matched onto the correct position.

4.5 Global Constraints (M3)

As mentioned in Section 4.3, the local matching procedure can easily confuse musically similar parts. Also, the computed segments obtained by individual matches may not be disjoint. In the procedure M3, we impose additional global constraints on the overall segmentation to cope with these two problems.

α_i	No.	$occ.$	$dur.$	S1	S2	M1	M2	M3	M4
1	0	23	216.5	0.995	0.994	0.968	0.975	0.975	0.975
2	0	23	283.3	0.996	0.995	0.977	0.976	0.976	0.976
3	0	23	081.4	0.962	0.972	0.881	0.918	0.918	0.918
4	1	23	069.9	0.888	0.927	0.900	0.937	0.937	0.937
5	1	22	070.9	0.808	0.938	0.753	0.747	**0.747**	**0.930**
6	1	23	138.4	0.826	0.986	0.969	0.970	0.952	0.952
7	2	23	122.9	0.854	0.983	0.930	0.932	0.932	0.932
8	2	23	152.4	0.959	0.992	0.977	0.977	0.977	0.977
9	2	23	139.8	0.987	0.987	0.986	0.988	0.970	0.977
10	3	22	073.1	0.930	0.945	0.775	0.772	**0.772**	**0.842**
11	3	23	230.3	0.989	0.992	0.985	0.985	0.985	0.985
12	3	23	074.6	0.990	0.993	0.964	0.967	0.967	0.967
13	3	23	092.1	0.939	0.982	0.979	0.976	0.976	0.976
14	4	23	034.6	0.749	0.876	0.617	**0.735**	**0.904**	0.904
15	4	23	029.3	0.635	0.798	0.496	**0.524**	**0.838**	0.838
16	4	20	026.4	0.550	0.692	0.519	**0.479**	**0.789**	0.789
17	5	23	186.0	0.979	0.985	0.930	0.930	0.930	0.930
18	6	23	287.8	0.984	0.994	0.987	0.989	0.989	0.989
19	7	23	223.9	0.963	0.972	0.992	0.992	0.992	0.992
20	8	23	499.4	0.989	0.997	0.995	0.994	0.994	0.994
21	9	23	258.6	0.979	0.992	0.945	0.988	0.988	0.988
22	9	23	137.6	0.971	0.978	0.985	0.980	0.980	0.980
23	10	22	337.3	0.944	0.951	0.944	0.943	0.987	0.987
24	10	23	301.9	0.977	0.986	0.989	0.988	0.981	0.981
25	10	23	243.8	0.910	0.986	0.933	0.932	0.924	0.924
26	10	23	059.7	0.740	0.889	0.908	0.847	0.883	0.883
27	11	19	104.5	0.631	0.725	0.807	0.807	**0.938**	0.938
28	12	23	356.9	0.882	0.988	0.982	0.982	0.982	0.982
29	13	22	161.5	0.794	0.940	0.943	0.943	0.986	0.986
30	13	22	208.8	0.814	0.951	0.945	0.944	0.984	0.987
31	14	23	168.4	0.729	0.923	0.796	0.790	**0.796**	**0.917**
32	14	19	**012.4**	0.439	0.643	**0.157**	**0.198**	**0.698**	0.735
33	14	22	057.7	0.714	0.846	0.864	0.869	0.827	0.913
34	15	23	147.2	0.745	0.946	0.938	0.937	0.980	0.980
35	16	23	303.6	0.827	0.996	0.990	0.989	0.989	0.989
36	16	23	503.2	0.812	0.965	0.994	0.994	0.994	0.994
37	16	23	241.2	0.781	0.894	0.987	0.987	0.987	0.987
38	16	23	068.2	**0.714**	**0.724**	**0.921**	**0.968**	0.968	0.968
∅		22.5	176.47	**0.852**	**0.930**	0.884	0.891	0.931	0.945

Table 2. Relative overlap values averaged over versions for different segments and different procedures. The first column (α_i) indicates the reference segment, the second column (No.) the musical number within the opera, the third column ($occ.$) the number of occurrences of α_i in the 23 versions of the dataset, and the fourth column ($dur.$) refers to the duration in seconds of α_i. S1, S2, M1, M2 M3, and M4 denote the respective segmentation procedures.

When using α_i as query, we now consider the entire ranked list of matches (instead of only using the top match as in M1 and M2). From each list we choose the best candidate so that the following global constraints are satisfied:

i) *Disjointness condition*: $\mathrm{P}(\alpha_i) \cap \mathrm{P}(\alpha_j) = \emptyset$

ii) *Temporal monotonicity*: $\alpha_i \prec \alpha_j \Rightarrow \mathrm{P}(\alpha_i) \prec \mathrm{P}(\alpha_j)$.

Here, we define the partial order $\prec$ on the set of segments by $\alpha_1 = [s_1 : t_1] \prec \alpha_2 = [s_2 : t_2] :\Leftrightarrow t_1 < s_2$. An optimal selection of matches from the ranked lists satisfying these global constraints can be computed using dynamic programming (similar to DTW). Howevever, note that in this case the dynamic programming is performed on the coarse segment level and not on the much finer frame level as in the case of global synchronization.

Applying this strategy does indeed improve the overall matching accuracy, on a version level as well as for individual segments, see Table 1 and Table 2. For example, for the segments α_{14}/α_{15}/α_{16}, the results improve from 0.735/0.524/0.479 for M2 to 0.904/0.838/0.789 for M3. Also, the results for α_{32} improve from 0.198 (M2) to 0.698 (M3).

Another interesting example is the relative overlap of 0.938 for α_{27}. This segment is actually missing in four

recordings of the opera. Using global constraints, the nonexistence of these segments was correctly identified by procedure $\mathrm{P} = \mathrm{M3}$ resulting in $\mathrm{P}(\alpha_{27}) = \emptyset$. However, the corresponding segment in `Leo1972` was misclassified as nonexistent by M3. A closer inspection revealed that the assumption modeled in the constraint that segments always appear in the same order as in the reference version was violated in this audio version. Here, the musical section covered by α_{27} was placed after α_{30} and used as an introduction before α_{31}. Thus, although strategy M3 stabilizes the overall matching, flexibility concerning the temporal order of segments is lost.

4.6 Structural Issues (M4)

Another problem occurs for the segments α_5, α_{10} and α_{31}, having the relative overlap values of 0.747, 0.772, and 0.796 for M3, respectively. According to the musical score, all these sections include repetitions of some music material. The segment α_5 for example should, according to the musical score, follow the structure $IA_1A_2B_1B_2O$, where I is an introductory and O an "outro" part. However, not all the repetitions are always performed. For example, the alternative structures IA_1B_1O, $IA_1A_2B_1O$, or $IA_1B_1B_2O$ for α_5 all appear in recordings of our dataset (similar variations occur for α_{10} and α_{31}). Such structural deviations can generally not be compensated well in the local matching procedure. Also, for further processing and analysis steps, such as the synchronization between corresponding segments in different recordings, it is important to know the exact structure of a given segment.

For M4, we investigate how structural correspondence of the query with an unknown version influences the segmentation quality. We manually annotated the musical structures occurring for α_5, α_{10} and α_{31} in the different audio versions of the opera. This information is then used in the matching to generate a query which structurally corresponds to the unknown version. The actual matching algorithm is the same as in M3. From the quantitative results in Table 2, we can conclude that the structural variations were indeed the cause of the poor performance for these segments: α_5 improves from 0.747 (M3) to 0.930 (M4), α_{10} from 0.772 (M3) to 0.842 (M4) and α_{31} from 0.796 (M3) to 0.917 (M4), see also Figure 3b.

5. CONCLUSIONS

In this paper, we presented a case study on segmenting given audio versions of an opera into musically meaningful sections that have been specified by a domain expert. Adapting existing synchronization and matching techniques, we discussed various challenges that occur when dealing with real-world scenarios due to the variability of acoustic and musical aspects. Rather than presenting technical details, our main motivation was to show how automated methods may be useful for systematically revealing and understanding the inconsistencies and variations hidden in the audio material. Furthermore, we showed how a procedure based on a combination of local matching and global constraints yields a more flexible and efficient alternative to a global black-box synchronization approach. Besides yielding slightly better results, this alternative procedure also provides a more explicit control to handle the various musical aspects and yields deeper insights into the properties of the audio material. For the future, we plan to expand our segmentation approach by explicitly including the dialogue sections into the analysis. Furthermore, the segmentation results will serve as basis for a finer grained analysis and multimodal processing including informed source separation.

Acknowledgments: This work has been supported by the BMBF project *Freischütz Digital* (Funding Code 01UG1239A to C). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS.

6. REFERENCES

[1] Simon Dixon and Gerhard Widmer. MATCH: A music alignment tool chest. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, GB, 2005.

[2] Emilia Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, UPF Barcelona, 2006.

[3] 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, NY, USA, 2003.

[4] Frank Kurth and Meinard Müller. Efficient index-based audio matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, 2008.

[5] 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)*, pages 1–6, 2011.

[6] Nicola Montecchio, Emanuele Di Buccio, and Nicola Orio. An efficient identification methodology for improved access to music heritage collections. *Journal of Multimedia*, 7(2):145–158, 2012.

[7] Meinard Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

[8] Meinard Müller and Sebastian Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 215–220, Miami, FL, USA, 2011.

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

[10] Meinard Müller, Henning Mattes, and Frank Kurth. An efficient multiscale approach to audio synchronization. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 192–197, Victoria, Canada, 2006.

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