Audio Engineering Society

Conference Paper

Presented at the Conference on
Semantic Audio
2017 June 22 – 24, Erlangen, Germany

A Multi-Version Approach for Transferring Measure Annotations Between Music Recordings

Frank Zalkow, Christof Weiß, Thomas Prätzlich, Vlora Arifi-Müller, and Meinard Müller

International Audio Laboratories Erlangen, Germany

Correspondence should be addressed to Frank Zalkow (`frank.zalkow@audiolabs-erlangen.de`)

ABSTRACT

In this paper, we address the task of transferring measure annotations between different recordings (versions) of a musical composition. Such annotations are useful for analyzing, linking, and navigating in multi-version scenarios of classical music. Given a version with manually annotated time positions, such as the beginning of musical measures, we transfer these annotations to musically corresponding positions in another version using synchronization techniques. As one contribution, we investigate the transfer process by exploiting additional versions. In a large-scale music scenario dealing with Richard Wagner's *Der Ring des Nibelungen*, we show that this multi-version analysis reveals musical passages that are problematic for synchronization. As another contribution, we introduce a late-fusion approach that improves the measure transfer when having several annotated versions.

1 Introduction

In Western classical music scenarios, one often has several music recordings of different performances (*versions*) of a musical composition. Sometimes, semantic annotations—such as structural segmentations, beat positions, or chord labels—are available for one version. In such cases, it might be useful to automatically transfer these annotations to other versions. For achieving this, alignments between the versions are necessary. Generating such alignments is the aim of music synchronization [1–5]. In this paper, we approach the automated transfer of measure annotations from one recording to another one using synchronization techniques. In the following, we assume that a *measure position* is specified by the time position of the beginning of the measure. Figure 1a illustrates how three measure positions are transferred from one version onto another one.

In real-world scenarios, it is often hard to evaluate the accuracy of the transferred annotations due to the lack of ground-truth annotations for the other versions. In this paper, we analyze the transferred annotations by means of a multi-version strategy introduced in [6]. This technique employs pairwise alignments for three versions. A time position in the first one is transferred to the musically corresponding time position in the second one, then from the second to the third one and finally, back to the first one. The comparison of the resulting time position with the initial one results in the *triple error*, which may indicate problems occurring during the synchronization procedure. Figure 1b shows an illustration of the triple error for three measure po-

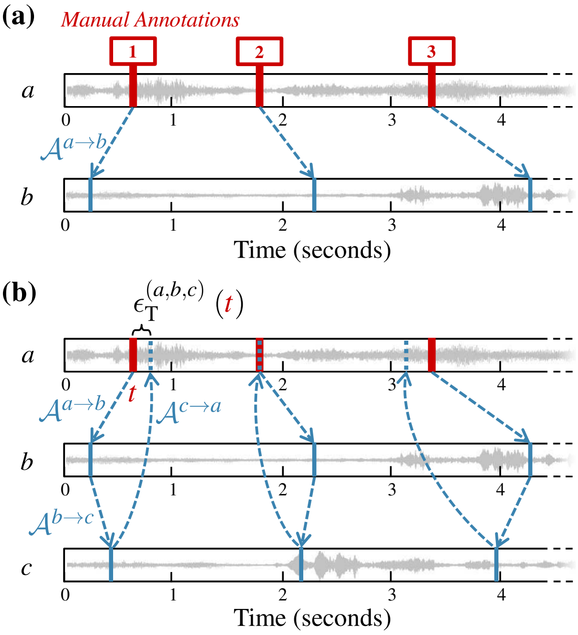

Fig. 1: Overview of the measure annotation transfer. **(a)** *Problem definition*: From a given reference version a with manual measure annotations, we compute an alignment $\mathcal{A}^{a\to b}$ to an unannotated version b and transfer the measure positions. **(b)** *Triple-based evaluation*: Using an additional version c, we compute alignments between all possible pairs. The comparison of the ground-truth positions with the triple-based transfer $a \to b \to c \to a$ yields the triple error $\epsilon_{\mathrm{T}}^{(a,b,c)}$.

sitions. Related multi-version approaches to improve alignment results were introduced in [7, 8].

In the following, we consider the cycle *Der Ring des Nibelungen* by Richard Wagner as a challenging case study. This work cycle comprises four music dramas with roughly 15 hours of music in total. In our experiments, we use a data set that contains six full performances of the *Ring* including both studio and live recordings. For three versions, we have manually generated measure annotations [9] and our goal is to transfer those to the unannotated versions. In such a transfer scenario, we call an annotated version a *reference version* and an unannotated one a *target version*.

As one contribution of this paper, we apply the triple-based evaluation procedure to evaluate the reliability of the transferred annotations in the *Ring* context. In particular, we identify specific situations where one typically observes high triple errors. Furthermore, we give musical and structural reasons leading to these errors.

As a second contribution, we propose a multi-version approach for transferring measure annotations to a target version. When having annotations for more than one version, it is not obvious which one to use as reference version for each measure. We propose to make this decision on the basis of the triple error. In our experiments, we compare our method to the straightforward approach of transferring all measures from a single reference version. Furthermore, we report on upper and lower bounds that could be achieved by this method. Our triple-based decision approach improves the accuracy compared to straightforward approaches.

The remainder of the paper is organized as follows. We first explain the triple error, describe our data set, and outline the synchronization method used in the experiments (Section 2). Then, we show how the triple error can be used to reveal musical passages that are problematic for automated synchronization procedures (Section 3). Finally, we introduce a new method for transferring annotations from several reference versions employing a late-fusion strategy (Section 4).

2 Prerequisites

In this section, we discuss prerequisites including a recapitulation of the triple error (Section 2.1), a description of our data set (Section 2.2), and an outline of the synchronization method used in the experiments (Section 2.3).

2.1 Triple Error

As a basis for our discussion, we recapitulate the definition of the pair and triple error as introduced in [6]. Let a and b be two different versions of a musical composition having durations $T_a, T_b \in \mathbb{R}_{>0}$ and time axes $[0,T_a]$ and $[0,T_b]$, respectively. An alignment $\mathcal{A}^{a\to b} : \mathbb{R} \to \mathbb{R}$ defines a mapping from the time axis $[0,T_a]$ of version a to the time axis $[0,T_b]$ of version b.

Let g_a denote the physical time point of a musical time position, for example, the beginning of a specific measure, in version a. g_b denotes the corresponding physical time position in version b. The pair (g_a, g_b)

is also denoted as ground-truth pair for the alignment. An errorless alignment $\mathcal{A}^{a\to b}$ maps the time position g_a to g_b. The *pairwise alignment error* ϵ_P indicates how close the automatically aligned position is to the ground-truth position:

$$\epsilon_\mathrm{P}^{(a,b)}\,(g_a) := \left|\mathcal{A}^{a\to b}(g_a) - g_b\right|. \qquad (1)$$

In the following, we assume that the ground-truth pairs relate to measure positions. In this case, $(g_a(k), g_b(k))$ constitutes an alignment between the time positions of the k^{th} measure position in versions a and b, respectively.

We are only able to compute the pairwise alignment error ϵ_P when ground-truth annotations are available. If this is not the case, there is another possibility for estimating the accuracy of the alignment. If we have at least three versions of the piece available, we can compute alignments for all pairs of versions. Starting from an arbitrary initial time position t in the first version, we can transfer it to the second version, then transfer the resulting time position to the third version and finally, transfer this resulting time position back to the first version. The difference between the resulting time position and the initial one is called the *triple error* ϵ_T, which is formally defined as

$$\epsilon_\mathrm{T}^{(a,b,c)}\,(t) := \left|\mathcal{A}^{c\to a}\Big(\mathcal{A}^{b\to c}\Big(\mathcal{A}^{a\to b}(t)\Big)\Big) - t\right| \qquad (2)$$

for the ordered triple (a,b,c), with $t \in [0, T_a]$. In Figure 1b, we show an illustration of three measure positions and their correspondng triple errors. We can interpret the triple error as an accumulation of errors produced during any of the involved alignments. In the case of all alignments being correct, the resulting triple error is necessarily zero. However, a triple error of zero is not a sufficient condition since positive and negative deviations in the alignment can cancel out, see [6] for a more detailed discussion. A high triple error is an indication for at least one involved alignment being erroneous.

When N versions are available, we can compute $N(N-1)(N-2)$ different ordered triples. When we want to visualize different triple errors on the same time axis, we have to keep the first version of the triple fixed. This yields $(N-1)(N-2)$ triple errors, which can be plotted on the same time axis. In the following, the triple error is always given in seconds.

ID	Part		Catalogue No.
A	*Das Rheingold*		WWV 86 A
B-1	*Die Walküre*	Act 1	WWV 86 B
B-2		Act 2	
B-3		Act 3	
C-1	*Siegfried*	Act 1	WWV 86 C
C-2		Act 2	
C-3		Act 3	
D-0	*Götterdämmerung*	Prologue	WWV 86 D
D-1		Act 1	
D-2		Act 2	
D-3		Act 3	

Table 1: Overview of Richard Wagner's cycle *Der Ring des Nibelungen.*

No.	ID	Conductor	Recording	hh:mm:ss
V_1*	Bar	Barenboim	1991–92	14:54:55
V_2*	Hai	Haitink	1988–91	14:27:10
V_3*	Kar	Karajan	1967–70	14:58:08
V_4	Bod	Bodanzky/Leinsdorf	1936–41	12:32:20
V_5	Bou	Boulez	1980–81	13:44:38
V_6	Sol	Solti	1958–65	14:36:58

Table 2: Performances of the *Ring* used in this paper. The * sign marks the versions for which ground-truth measure annotations are available. In V_6, Leinsdorf only conducts *Die Walküre* while the other parts are conducted by Bodanzky.

2.2 Data Set

We consider the cycle *Der Ring des Nibelungen* by Richard Wagner comprising four music dramas with 21952 measure positions in total.[1] Table 1 shows an overview of the music dramas of the *Ring* and their subparts. Our data set consists of six complete recordings of the *Ring*, each lasting between 12 and 15 hours, resulting in 85 hours of music in total. Table 2 gives an overview of the versions. For three of these, manual measure annotations are available. They were created by students with a background in Western classical

[1] There are 21939 measures in the *Ring*, but we also include the pickup measure of two subparts and the end of the final measure of each subpart as positions.

music and passed several revision cycles involving multiple annotators for each measure [9]. We use these annotations as ground-truth for our experiments. In the plots, we use the performance number (Table 2) to indicate the triples. A concatenation of work ID (Table 1) and performance ID (Table 2) refers to a specific piece in a specific version, for example, `C-2/Sol` denotes to the second act of *Siegfried*, conducted by Solti.

2.3 Synchronization Method

A music synchronization algorithm generates an alignment of musically corresponding time positions between two or more recordings of a musical piece [10, Chapter 3]. This paper evaluates and partially improves the alignment while considering the synchronization algorithm as a black box. Therefore, our findings are applicable for any synchronization method such as [1–5, 11, 12]. For clarity, we summarize the synchronization pipeline used in our experiments.

We use a chroma-based synchronization procedure, which additionally incorporates onset-based features to improve the accuracy, see [3] for details. Due to runtime issues, a multi-scale dynamic time warping procedure is used [13]. The feature rate is 50 Hz, thus resulting in a temporal resolution of 20 ms. This leads to a discretization of the alignment, which is also reflected in our results. Nevertheless, using suitable interpolation techniques, the discrete alignment can be converted to continuous time.

3 Detecting Synchronization Problems

We now demonstrate how the triple error can be used to analyze the alignment in our music scenario. In particular, we identify situations where we observe high triple errors that indicate alignment problems. Furthermore, we give musical reasons leading to these errors. We illustrate three typical situations where music synchronization is problematic comprising the beginning and closing of a piece of music (Section 3.1), structurally differing performances such as abridged versions (Section 3.2), as well as passages with a high degree of homogeneity with regard to a musical aspect relating to the feature representation used (Section 3.3).

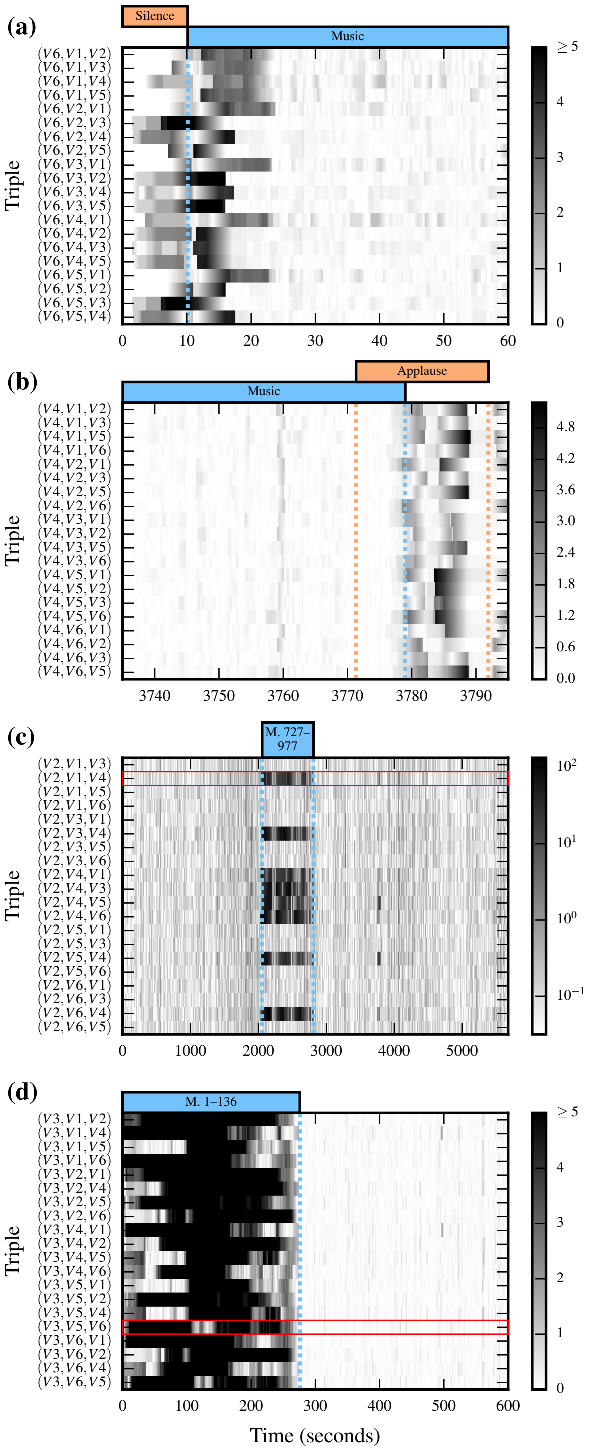

Fig. 2: Triple errors for several sections with alignment problems. **(a)** Beginning of `C-2/Sol` (V_6). **(b)** Ending of `C-3/Bod` (V_4). **(c)** Complete `B-2/Hai` (V_2). **(d)** Beginning of `A/Kar` (V_3).

3.1 Beginning and Closing Sections

Figure 2 shows the triple error for several triples (vertical axis) across time (horizontal axis). In our plots, darker colors correspond to a higher triple error. The boxes above the plots mark ground-truth annotations of musical structure relating to the behaviour of the triple error. In Figure 2a, we can observe that many triples lead to high errors in the first 20 seconds, which corresponds to the beginning of `C-2`. We can interpret this as an indication for an unreliable alignment at the beginning. Recordings can have silent passages of different length preceding and succeeding the actual music. In those sections, synchronization procedures generate more or less random alignments. Furthermore, at the beginning and ending of the actual music, some time is needed before the synchronization procedures stabilizes to produce accurate alignments. Therefore, an alignment often exhibits errors in those sections. In particular, this applies to musical passages like the beginning of `C-2`, which opens very softly.

In Figure 2b, we see a similar situation between seconds 3780–3790 corresponding to the closing of `C-3`. Besides the silence problem at the end, this example is a live recording with applause. Sometimes, an enthusiastic audience begins clapping before the end of the piece, which leads to an overlay of the recorded music with applause. Interestingly, we observed for `C-3/Bod` that the triple error increases only when the music ends. Since the triple error is lower in the section where the music is overlaid with applause, it seems reasonable to conclude that the applause does not have a strong negative effect on the alignment in this example.

3.2 Structural Differences

In Figure 2c, we can see another interesting example with high triple errors. When starting from `B-2/Hai` (V_2), there is a section with very high triple errors between seconds 2000–3000. Opposed to the previous two examples, the triple error is high only for triples that involve `B-2/Bod` (V_4). This points to an abridgement in this version, where we find omissions between measures 727–977. To clarify this point, let us consider the triple (V_2, V_1, V_4) as an example, and measure 810 as a starting point. This measure can be aligned well from V_2 to V_1. But the alignment from V_1 to V_4 is meaningless since this measure is omitted in V_4. The alignment accidentally leads to measure 782 in V_4, which is not omitted there. Finally, measure 782 can be aligned well to V_2. Accordingly, the comparison of the starting point at measure 810 and the resulting measure 782 yields a high triple error. In Wagner's works, the music of an act has usually neither interruptions nor exact repetitions (*through-composed*). Therefore, we did not expect abridgements inside an act. However, it was common practice to cut Wagner's works at the Metropolitan Opera during the 1910s–40s, since some conductors thought that the cuts would prevent the audience from becoming bored [14, Chapter 14].

Aligning versions with structural differences—for example, abridged versions—is a major challenge in automated music synchronization. There is no straightforward method to align structurally different sections and this problem already has got some attention. A method for transferring a segment annotation to an abridged version of the same piece can be found in [15]. Alignment techniques for versions with structural differences are proposed in [12, 16, 17].

3.3 Homogeneous Passages

As a final example, Figure 2d reveals very high triple errors in the first 250 seconds. This section corresponds to the prelude of *Das Rheingold*, which is a very homogeneous passage of music with respect to harmony. It exhibits 136 measures with a constant harmony, an E♭ major triad. When the alignment is based on a feature representation that mainly relates to harmony—such as variants of chroma features—the alignment is not stable in such situations. This is a well known challenge in music synchronization [9].

4 Improving Measure Transfers

As a further contribution of this paper, we present a new method for transferring annotations from several reference versions employing a late-fusion approach (Section 4.1), which is then evaluated in the *Ring* context (Section 4.2).

4.1 Fusion Strategy

As said before, an important application of music synchronization is the transfer of annotations between different representations of a musical composition [2]. In a score-to-audio alignment scenario, such measure transfers were approached in [9]. In this section, we want to transfer measure annotations between different audio versions. In the case of several reference versions, we show how the triple error can be used to improve the measure transfer to the target version in a late-fusion approach. This is related to multi-version alignment approaches as discussed in [7, 8].

In Figure 2d, we can see that some passages exhibit a much smaller triple error in specific triple constellations than in others. At 100–150 seconds, for example, the triple error for (V_3,V_5,V_6) is much smaller than for other triples. Locally, some alignments seem to work better than others. Our idea is to exploit the best local alignment for transferring measure annotations. To this end, we propose to use the triple error for improving the results of a measure transfer task using a late-fusion strategy. When we have several reference versions and we want to transfer annotations to a target version, we select the reference version with the best local alignment for each measure. The errors for triples that include a specific alignment give us an indication for the accuracy of this alignment.

In the following, we assume that we have $N > 3$ versions of a musical piece. In our experiments, we use $N = 6$ versions. Furthermore, we assume that we have annotations for the first two versions $n \in \{1,2\}$. Let $g_1(k)$ and $g_2(k)$ be the time position of the k^{th} measure position in reference version 1 and 2, respectively. We want to find out the corresponding time position in target version 3 and can choose to transfer it either from reference version 1 or 2. The better one of the two is called $n^*(k) \in \{1,2\}$. The transfer from $n^*(k)$ gives us an estimate for the time position on the target version:

$$\widetilde{g}_3(k) = \mathcal{A}^{n^*(k)\to 3}\left(g_{n^*(k)}(k)\right). \qquad (3)$$

To select $n^*(k)$, we estimate the accuracy for both alignments that transfer the measure position k from one of the reference versions $n \in \{1,2\}$ to target version 3. To this end, we compute triple errors involving the additional versions. An estimate for the better version of the two is found by computing the triple error for triples that start with a fixed reference version n and transfer it to the target version 3. The third version of the triple can be any version $m \in \{4,\ldots,N\}$. For each $n \in \{1,2\}$, we can compute $N-3$ triple errors that indicate the accuracy of the transfer from n to 3. The version with the least median[2] triple error is chosen to be $n^*(k)$. This is done separately for each measure position k:

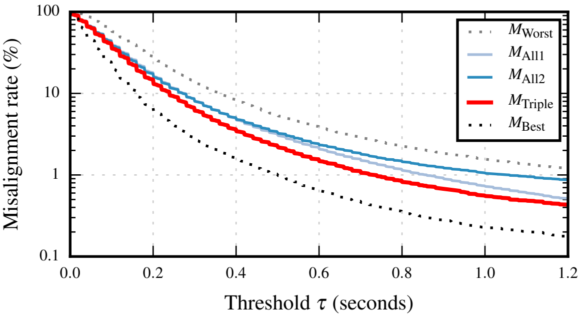

Fig. 3: Misalignment rates for straightforward approaches, proposed method and oracle-fusions.

$$n^*(k) = \underset{n\in\{1,2\}}{\operatorname{argmin}}\left(\underset{m\in\{4,\ldots,N\}}{\operatorname{median}}\left(\epsilon_{\mathrm{T}}^{(n,3,m)}\left(g_n(k)\right)\right)\right). \qquad (4)$$

Without loss of generality, we introduced the method with a fixed number of two reference versions. The method can easily be generalized to more reference versions.

4.2 Evaluation

For evaluating our approach, we transfer measure annotations from reference versions 1 and 2 to a target version 3 with different methods. In our experiments, we also have ground truth annotations g_3 for version 3 to compute evaluation scores. In the following, we refer to the proposed triple-fusion method as M_{Triple}. We compare this strategy with a straightforward method of transferring all measures from only one of the reference versions, called M_{All1} and M_{All2}, respectively. Furthermore, we introduce two "oracle-fusion" methods referred to as M_{Best} and M_{Worst}: For each measure, we transfer the measure position from the recording (reference version 1 or 2) that leads to the lowest or highest pair error, respectively.

For the following experiment, we transfer all 21952 measure positions of the *Ring* three times separately. In the three cases, we use (V_2,V_3), (V_1,V_3) and (V_1,V_2) as reference versions, where the first element of each tuple corresponds to version 1 and the second one corresponds to version 2. The target version is V_1, V_2 and V_3, respectively. Figure 3 shows the misalignment rates averaged over all three transfer cases. The misalignment rate is the percentage of aligned measure positions

[2] Experiments showed that mean and median behave rather similar for our method. Note, for $N = 4$ we only have a single additional version for selecting $n^*(k)$. That leads to taking the median of a single triple error and this means taking the value itself or, in other words, taking no median at all. For $N = 5$ we have two additional versions and the median is equal to the mean.

	μ	σ	P_{25}	P_{50}	P_{75}	P_{85}	P_{95}	P_{98}	P_{99}	$P_{99.25}$	$P_{99.5}$	$P_{99.75}$	$P_{99.9}$
M_{All1}	134	525	40	80	160	220	399	620	860	984	1207	1636	2545
M_{All2}	171	812	40	80	154	217	400	664	1060	1440	2340	8501	16617
M_{Triple}	128	609	40	80	140	194	340	525	732	840	1060	1861	5058
M_{Worst}	230	953	70	120	211	283	520	860	1410	1760	2992	9020	16652
M_{Best}	75	122	20	40	92	127	220	360	500	560	675	952	1520

Table 3: Statistics on the pair error using different methods, in milliseconds (rounded). P_i refers to the i^{th} percentile.

whose pair error is above a specified threshold $\tau \in \mathbb{R}_{>0}$ in seconds, see also [2, 6, 7]. We see that, on average, our triple-based method outperforms the transfer procedures M_{All1} and M_{All2}. For example, about 2.2 % (M_{Triple}) of the measures have an error greater than $\tau = 0.5$ seconds, compared to 3.1 % (M_{All1}) and 3.3 % (M_{All2}). It is a non-trivial result to outperform M_{All1} and M_{All2} since in practice we have no ground-truth for the target version. Thus, there is no indication if M_{All1} or M_{All2} is the better choice.

To obtain deeper insights into these results, let us consider some further statistics. In Table 3, we show the mean μ, the standard deviation σ, and the i^{th} percentile P_i (for some selected i) over all pair errors. We see that the improvements of our method particularly apply to about 5 % of the measures with the highest pair error. The percentile P_{99}, for example, represents the pair error below which 99 % of the measures fall—or, in other words, it shows the lowest pair error of the 1 % most erroneous measures. This value is 732 ms with the triple-based approach (M_{Triple}), compared to 860 ms (M_{All1}) and 1060 ms (M_{All2}). A lower and an upper bound are given by 1410 ms (M_{Worst}) and 500 ms (M_{Best}), respectively. The majority of measures are not affected. The value for P_{50} indicates that 50 % of the measures have a pair error below 80 ms, no matter if a straightforward approach or our method is used. Even if our improvements affect only a small fraction of the measures, these effects can be of practical relevance in large scale scenarios. For our *Ring* data set, for example, 1 % corresponds to roughly 220 measures.

To evaluate how sensitive our method is to the number of additional versions used for the voting strategy, we performed an experiment with varying amount of versions (Table 4). We have three possibilities for selecting one (or two) additional versions from our data set of six versions ($N = 4$ and $N = 5$). Therefore, we performed our method three times for each target version and report the statistics over all computed pair errors. For $N = 18$, we included twelve further versions to our data set. P_{99} drops from 761 ms to 695 ms for an increase of N from 4 to 18, whereas the average pair error is quite stable. We see that our method is valuable even for a single additional version and there is no major difference for a varying amount of versions. One reason for this is that in the triple-based approach the additional versions are used indirectly, only for voting, and not directly as, for example, in multi-version alignment approaches [7, 8].

	μ	P_{85}	P_{95}	P_{99}
$N = 4$	135	200	354	761
$N = 5$	133	200	340	750
$N = 6$	128	194	340	732
$N = 18$	131	180	320	695

Table 4: Statistics on the pair error for M_{Triple} with varying amount of $N-3$ additional versions for voting, in milliseconds (rounded).

5 Summary

In this paper, we approached the automated transfer of measure positions between different recordings of a musical composition by means of synchronization techniques. We presented a multi-version approach for analyzing the computed alignments and thus, estimated the accuracy of the transfer. We showed that the *triple error* can be used to identify sections where the alignment is problematic and provided musical reasons for meaningful examples. Furthermore, we proposed an approach for exploiting additional versions to improve the accuracy of the transfer from several annotated versions. We evaluated this method in a large-scale music scenario comprising six performances of Richard Wagner's cycle *Der Ring des Nibelungen*. These experiments showed that the use of additional versions is beneficial for the annotation accuracy. A large amount of additional versions used for voting only slightly increases the effectiveness of our method. Actually, a single additional version is sufficient for improving the measure transfer.

6 Acknowledgments

We thank all students involved in the annotation work, in particular Cäcilia Marxer. This work was supported by the German Research Foundation (DFG MU 2686/7-1, DFG MU 2686/6-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

References

[1] Orio, N. and Schwarz, D., "Alignment of Monophonic and Polyphonic Music to a Score," in *Proc. of the Int. Computer Music Conf. (ICMC)*, pp. 155–158, Havana, Cuba, 2001.

[2] Dixon, S. and Widmer, G., "MATCH: A Music Alignment Tool Chest," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 492–497, London, UK, 2005.

[3] Ewert, S., Müller, M., and Grosche, P., "High Resolution Audio Synchronization Using Chroma Onset Features," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1869–1872, Taipei, Taiwan, 2009.

[4] Arifi, V., Clausen, M., Kurth, F., and Müller, M., "Synchronization of Music Data in Score-, MIDI- and PCM-Format," volume 13 of *Computing in Musicology*, pp. 9–33, MIT Press, 2004.

[5] Montecchio, N. and Cont, A., "A Unified Approach to Real Time Audio-to-Score and Audio-to-Audio Alignment using Sequential Montecarlo Inference Techniques," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 193–196, Prague, Czech Republic, 2011.

[6] Prätzlich, T. and Müller, M., "Triple-Based Analysis of Music Alignments Without the Need of Ground-Truth Annotations," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 266–270, Shanghai, China, 2016.

[7] Arzt, A. and Widmer, G., "Real-Time Music Tracking Using Multiple Performances as a Reference," in *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, pp. 357–363, Málaga, Spain, 2015.

[8] Wang, S., Ewert, S., and Dixon, S., "Robust and Efficient Joint Alignment of Multiple Musical Performances," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11), pp. 2132–2145, 2016.

[9] Weiß, C., Arifi-Müller, V., Prätzlich, T., Kleinertz, R., and Müller, M., "Analyzing Measure Annotations for Western Classical Music Recordings," in *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, pp. 517–523, New York, USA, 2016.

[10] Müller, M., *Fundamentals of Music Processing*, Springer Verlag, 2015.

[11] Dannenberg, R. B. and Hu, N., "Polyphonic Audio Matching for Score Following and Intelligent Audio Editors," in *Proc. of the Int. Computer Music Conf. (ICMC)*, pp. 27–34, San Francisco, USA, 2003.

[12] Joder, C., Essid, S., and Richard, G., "A Conditional Random Field Framework for Robust and Scalable Audio-to-Score Matching," *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8), pp. 2385–2397, 2011.

[13] Prätzlich, T., Driedger, J., and Müller, M., "Memory-Restricted Multiscale Dynamic Time Warping," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 569–573, Shanghai, China, 2016.

[14] Brown, J., *Great Wagner Conductors: A Listener's Companion*, Parrot Press, 2012.

[15] Prätzlich, T. and Müller, M., "Frame-Level Audio Segmentation for Abridged Musical Works," in *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, pp. 307–312, Taipei, Taiwan, 2014.

[16] Müller, M. and Appelt, D., "Path-Constrained Partial Music Synchronization," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pp. 65–68, Las Vegas, USA, 2008.

[17] Grachten, M., Gasser, M., Arzt, A., and Widmer, G., "Automatic Alignment of Music Performances with Structural Differences," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 607–612, Curitiba, Brazil, 2013.