TRIPLE-BASED ANALYSIS OF MUSIC ALIGNMENTS WITHOUT THE NEED OF GROUND-TRUTH ANNOTATIONS

Thomas Prätzlich and Meinard Müller

International Audio Laboratories Erlangen

ABSTRACT

The goal of music alignment methods is to temporally align different versions of the same piece of music. These methods are typically evaluated by comparing the computed alignments to given ground-truth annotations. Creating such annotations is usually very labor intensive. For many musical pieces, especially in classical music, there exists a multitude of different recordings. In this work, we investigate whether an evaluation of music alignment algorithms can be performed without ground-truth annotations when at least a triplet of recordings of the same piece of music is available. The main idea is to align the time points of a fixed reference version, in a circular way, back through a second and third version by using their pairwise alignments. A triple error is then computed by comparing these time points with their circularly aligned version. In this paper, we formalize the idea of the triple error and discuss its potential and limitations. We present typical examples for the triple error and compare it to the pairwise alignment error based on ground-truth. Furthermore, we present a case study to indicate the potential of the triple error to analyze alignments and to compare different alignment methods without the need of ground-truth annotations.

Index Terms— music alignment, music synchronization, evaluation, ground-truth

1. INTRODUCTION

In the field of Music Information Retrieval, music alignment algorithms are an important tool and active area of research [1, 2, 3, 4]. The goal of music alignment is to find corresponding time positions between different versions of the same piece of music. In an *audio-audio* alignment scenario, the task could be to align two different versions of the Mazurka Op. 63 No. 3 composed by Chopin, one performed by Rubinstein and the other by Cohen. To evaluate the quality of such an alignment, ground-truth annotations marking corresponding time positions in the two recordings are needed. These time positions are typically note onsets, beats, or measure positions. However, ground-truth annotations are not always available and their manual creation is very labor intensive. Furthermore, an evaluation is restricted to the time positions of the ground-truth annotations.

Often, especially in classical music, more than two versions of the same piece of music are available, sometimes even in different representations. For example, different audio recordings and a musical score representation might be available. More recently, multiple versions available for the same piece of music were used jointly, to improve and stabilize alignment methods [3, 4, 5, 6].

The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the Fraunhofer-Institut für Integrierte Schaltungen (IIS). This work has been supported by the BMBF project *Freischütz Digital* (Funding Code 01UG1239A to C) and the German Research Foundation (DFG MU 2686/7-1)

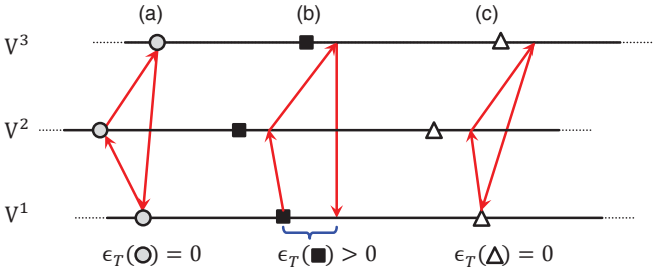

Fig. 1. Illustration of the concept of a triple error on the time axis of three different versions V_1, V_2, and V_3. The red arrows indicate alignments between the different versions. The circles, squares and triangles mark corresponding ground-truth positions in the different versions. The triple error is denoted by ϵ_T. **(a)** Consistent alignment with triple error of zero. **(b)** Inconsistent alignments with a triple error greater zero. **(c)** Inconsistent alignments with zero triple error.

In this work, we exploit the availability of different versions for a ground-truth independent evaluation and analysis of alignment methods. For example, when considering a third recording of the Mazurka Op. 63 No. 3 performed by Ezaki, we can build a triplet of recordings of the same piece of music. Our main idea is to use the pairwise alignments between the versions in such a triplet in a circular way, to evaluate the alignments without the need for ground-truth annotations. An arbitrary given starting point on the time axis of a version V_1 is then circularly aligned back onto the time axis of V_1 through a version V_2 and a version V_3. In this way, a triple error can be computed by measuring the difference between the starting point and its circularly aligned version (see Figure 1). In the case of alignments that consistently align corresponding time positions in the different versions, we expect the triple error to be zero (see Figure 1a). In the presence of alignment errors, we expect the triple error to be greater than zero (see Figure 1b). However, note that the triple error might be higher or lower than a pairwise alignment error measured with ground-truth annotations. Errors from the different alignments can either accumulate or cancel out. The latter is illustrated in Figure 1c, where the triple error is zero, despite the fact that there are errors in the pairwise alignments. Hence, measuring a zero triple error is only a necessary condition for the alignment to be correct. However, a triple error greater than zero still implies errors in at least one of the alignments involved in the triple error computation.

In this paper, we formalize the concept of the triple error (Section 2) and discuss its theoretical limitations. Then we perform experiments to illustrate that in practice, despite its theoretical limitations, the triple error can be a useful tool for analyzing and evaluating alignment results (Section 3). Related work will be discussed in the respective sections.

ID	Composer	Piece	$\overline{\text{dur}}[s]$	Type	#(GT)	#(GT)/s
F06	Weber	Freischütz, No. 6	297	measures	150	0.51
F08	Weber	Freischütz, No. 8	540	measures	199	0.37
F09	Weber	Freischütz, No. 9	428	measures	200	0.47
M17-4	Chopin	Op. 17, No. 4	247	beats	396	1.60
M24-2	Chopin	Op. 24, No. 2	124	beats	360	2.90
M63-3	Chopin	Op. 63, No. 3	144	beats	229	1.59
M68-3	Chopin	Op. 68, No. 3	89	beats	181	2.03

Table 1. The three numbers of "Der Freischütz" and the four Chopin Mazurkas used in our experiments. For each piece, there are three recordings in our dataset with an average duration of $\overline{\text{dur}}[s]$. **#(GT):** number of annotated ground-truth positions, **Type:** type of annotations (measure or beat annotations), **#(GT)/s:** average number of ground-truth annotations per second.

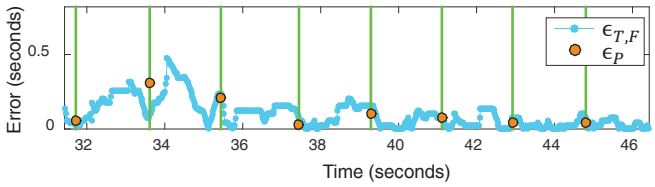

Fig. 2. Example of frame-wise triple error $\epsilon_{T,F}$ (cyan curve) and pairwise alignment error ϵ_P (orange dots) on the piece F06 using the performance by Kleiber as reference. The solid green lines show the ground-truth grid.

2. TRIPLE ERROR

Let V_1 and V_2 be two versions of the same piece of music with the corresponding time-continuous axes $[0, T_1]$ and $[0, T_2]$. An *alignment* $\mathcal{A}$ from V_1 to V_2 is a mapping of time points from V_1 onto corresponding time points of V_2. We model an alignment by the function

$$\mathcal{A}\colon [0, T_1] \to [0, T_2] \tag{1}$$

which is monotonous, i.e. $\mathcal{A}(s) \leq \mathcal{A}(t)$ for $s, t \in [0, T_1]$ with $s \leq t$. Furthermore, an alignment fulfills the boundary constraints $\mathcal{A}(0) = 0$ and $\mathcal{A}(T_1) = T_2$. Sometimes further constraints on the slope are required. Typically, an alignment is evaluated using pairs of manually specified ground-truth annotations that mark corresponding time positions in the two different versions. Each ground-truth pair is specified as

$$(g_1, g_2) \in [0, T_1] \times [0, T_2]. \tag{2}$$

Using these ground-truth pairs, the *pairwise alignment error* for a given alignment $\mathcal{A}$ between two versions is defined as $\epsilon_P\colon [0, T_1] \to \mathbb{R}$ with

$$\epsilon_P(g_1) := |\mathcal{A}(g_1) - g_2|\,. \tag{3}$$

The main drawback of this error measure is the need for ground-truth annotations which are often unavailable.

We now formalize our main idea, which exploits the fact that for many pieces, more than two versions are available. Let the triplet (V_1, V_2, V_3) be three versions with corresponding time axes $[0, T_i]$, where $i \in \{1, 2, 3\}$. Furthermore, let V_1 serve as a reference version. Assume the alignments $\mathcal{A}^{1\to2}, \mathcal{A}^{2\to3}, \mathcal{A}^{3\to1}$ are given, where $\mathcal{A}^{i\to j}$ denotes the alignment of version V_i to V_j. Using these alignments in a circular way, a time point $t \in [0, T_1]$ of version V_1 can be subsequently aligned to V_2, V_3 and back to V_1 by applying the alignments in a composition to compute $t' := \mathcal{A}^{3\to1}\left(\mathcal{A}^{2\to3}\left(\mathcal{A}^{1\to2}(t)\right)\right) \in [0, T_1]$. We can now measure the difference between the original time point t and its circularly aligned version t' by the *triple error* $\epsilon_T\colon [0, T_1] \to \mathbb{R}$ that is computed by

$$\epsilon_T := \left|\mathcal{A}^{3\to1}\left(\mathcal{A}^{2\to3}\left(\mathcal{A}^{1\to2}(t)\right)\right) - t\right|. \tag{4}$$

The main advantage of the triple error compared to the pairwise alignment error is its independence of ground-truth annotations. However, from a theoretical point of view, the triple error has to be considered with great care. First, note that the triple error is based on three pairwise alignments. Hence, when measuring a triple error we know that the pairwise alignment error in one of the alignments is at least one third of the measured triple error. However, the exact pairwise alignment errors cannot be inferred from the triple error. Second, a zero triple error does not necessarily imply that the alignments are error free. An error introduced by one alignment can be compensated by another (see Figure 1c). There are even alignments that always lead to a zero triple error. We can construct such an alignment by linearly scaling the durations of the different versions to be aligned. This leads to a triple error of zero for all time points, although the pairwise alignment error might be high. This proves that *a zero triple error is only a necessary condition* for a zero pairwise alignment error, but not a sufficient one. However, measuring a *triple error greater than zero is a sufficient condition* that there is an error in one of the pairwise alignments involved in the triplet. Hence, the triple error can still be a useful measure for the analysis of alignments. In practice, when using reasonable alignments, we expect the triple error to reflect the values of the pairwise alignment error in most of the cases.

3. EXPERIMENTS

In this section, we first review the dataset (Section 3.1) and the alignment approach (Section 3.2) used in our experiments. Within the remaining sections, we perform experiments to compare the triple error with the pairwise alignment error. These experiments indicate that often, the same conclusions as drawn from a ground-truth based analysis can be made, by only using the ground-truth independent triple error. In particular, we discuss the potential of the triple error to identify problematic positions within an alignment (Section 3.3). Furthermore, we show how the triple error can be used to identify problematic versions (Section 3.4). Finally, we indicate its potential to compare different alignment methods (Section 3.5).

3.1. Dataset

For our experiments, we use a set of three excerpts of the opera "Der Freischütz" and four Chopin Mazurkas [7] (see Table 1). We use measure annotations (marking musical measure boundaries) for the opera excerpts and beat annotations for the Chopin Mazurkas (marking beat positions). For a given version, we denote all available ground-truth time positions as *ground-truth grid*. The Mazurkas are piano pieces that typically have clear note onsets, whereas the operas are composed of singing and orchestral music which typically has soft or blurred onsets. Using pieces with different instrumentations allows us to identify problems of an alignment method related to specific instrumentations. For each piece, we consider a triplet of recordings. For the opera excerpts the recordings were conducted by Bloemeke (2013), Kleiber (1973), and Furtwaengler (1954). The Mazurka recordings were performed by Ezaki (2006), Cohen (1997), and Rubinstein (1966). In our dataset, there is no canonical order of the different versions. We therefore compute all pairwise alignments

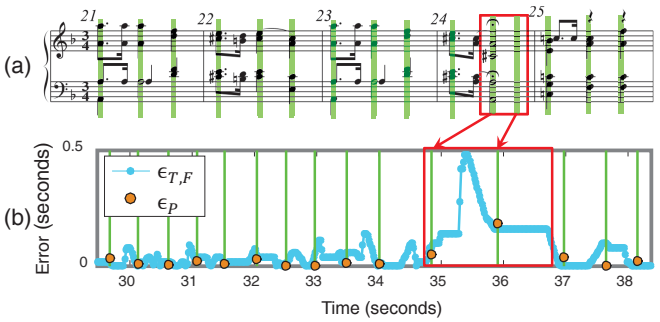

Fig. 3. Excerpt showing frame-wise triple error $\epsilon_{T,F}$ and pairwise alignment error (solid cyan curve) ϵ_P (orange dots) on the piece M68-3 using the performance by Cohen as reference. The green lines mark the ground-truth grid. **(a)** Musical score excerpt of M68-3. **(b)** frame-wise triple error corresponding to the score excerpt shown in (a).

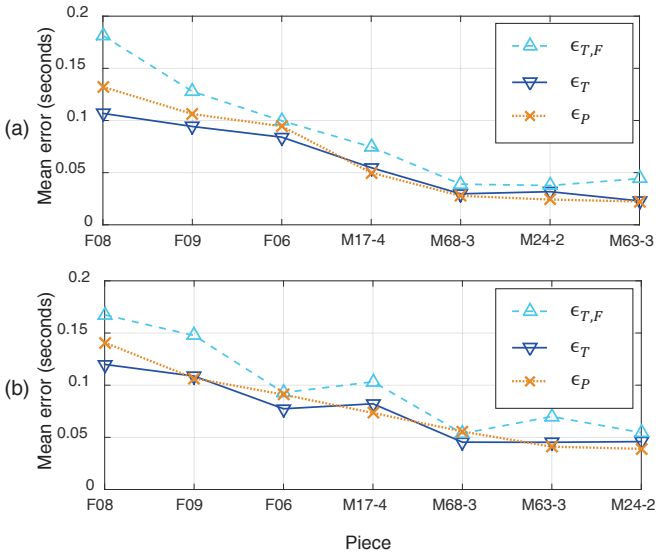

Fig. 4. Mean values for frame-wise triple error $\epsilon_{T,F}$, triple error ϵ_T (computed on ground-truth grid), and pairwise alignment error ϵ_P for the pieces in our dataset. The pieces are sorted in descending order with respect to the mean pairwise alignment error. **(a)** Alignment method $\mathcal{A}_1$ with onset informations, **(b)** Alignment method $\mathcal{A}_2$ using only chroma features without onset information.

$\mathcal{A}^{i\to j}$ with $(i,j) \in [1:3]^2$ and $i \neq j$. This leads to six possible pairs for which we compute six triple errors (two for each version as a reference) and six pairwise alignment errors.

3.2. Alignment Method

The objective of music alignment is to compare two given sequences corresponding to different versions of the same piece of music. Many different alignment methods have been proposed, e.g. [8, 9, 10, 2, 11, 12]. In the following experiments, we use an alignment procedure based on Dynamic Time Warping (DTW) [13]. The goal of DTW is to find an optimal alignment of two sequences under certain restrictions. In the case of two different versions of the same piece of music, one first needs to compute a suitable feature representation. Chroma features capture the coarse harmonic progression of the music and are commonly used in music alignments [2, 13, 14, 11]. To achieve a high temporal accuracy, we use the alignment approach introduced in [14], where chroma-based features are combined with features capturing note onset informations. Furthermore, we extend the approach to only align two given versions between the first and last ground-truth annotation for each recording. In this way, we exclude boundary artifacts that are caused by aligning silence or non-musical sounds such as applause or noise. All alignments were computed at a feature rate of 50 Hz.

Note that in Section 2, an alignment was modeled on a continuous time axis. The computed alignments, however, are time-discrete. A time-discrete alignment is a sequence of pairs containing corresponding time positions of the aligned versions. This makes it necessary to use interpolation to align arbitrary time positions across different versions [15].

In the following, we mainly discuss three error measures. First, the pairwise alignment error ϵ_P that is computed on the ground-truth grid using Equation (3). Second, the triple error ϵ_T that is evaluated only at the time positions of the ground-truth grid. And third, the frame-wise triple error $\epsilon_{T,F}$ that is evaluated on a 50 Hz frame grid which corresponds to the feature resolution of the alignments. The two last measures are computed with Equation (4).

3.3. Identification of Alignment Errors

Figure 2 shows an example of the frame-wise triple error $\epsilon_{T,F}$ and the pairwise alignment error ϵ_P on an excerpt of the piece F06. The green solid lines indicate the ground-truth grid of musical measures at which ϵ_P is computed. In this example, we can clearly see that the pairwise alignment error and the triple error show mostly the same tendencies when comparing them at the positions of the ground-truth grid. Note that the ground-truth grid marks characteristic positions in the recordings that usually coincide with note onsets. At these positions, the alignment is more likely to have lower errors. The positions in between the grid positions often correspond to steady notes. These lead to regions of homogenous feature values where the alignment typically exhibits higher error values.

Figure 3a+b shows an excerpt of the piece M68-3 with the corresponding musical score together with the pairwise alignment error ϵ_P and the frame-wise triple error $\epsilon_{T,F}$. Here again, the two error measures coincide well at the ground-truth grid, which marks beat positions in this example. For the same reason as above, the triple error is again much larger in between the positions of the ground-truth grid. Note that the beat annotations not always go along with onset positions and vice versa. There are *silent beats* that have no associated onset. For example, in Figure 3, the second beat of the half note within the red rectangles is silent. At the position of the silent beat, both the pairwise alignment error and the triple error exhibit high values. This illustrates that the higher errors between note onsets are not an artifact of the frame-wise triple error. Hence, the triple error can sometimes even lead to additional insights.

Figure 4 shows the mean of the triple errors $\epsilon_{T,F}$ and ϵ_T, and the mean of the pairwise alignment errors ϵ_P for all pieces in the dataset, see also Table 1. Note that the visualization of the results is sorted by decreasing values of the pairwise alignment error. Overall, the triple errors and the pairwise alignment errors show similar tendencies. Especially the triple error ϵ_T that is evaluated on the ground-truth grid is very close to the pairwise alignment error in most cases. Furthermore, in Figure 4a, note that the frame-wise triple error $\epsilon_{T,F}$ is always higher than the triple error on the ground-truth grid ϵ_T caused

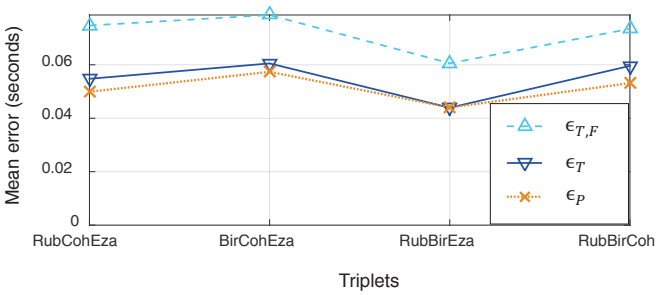

Fig. 5. Mean of frame-wise triple error $\epsilon_{T,F}$, the triple error ϵ_T computed on the ground-truth grid and mean pairwise alignment error ϵ_P on M17-4. The labels on the x-axis denote which recordings are used in a triplet, using the recordings by Biret (Bir), Cohen (Coh), Ezaki (Eza), and Rubinstein (Rub).

by the larger errors in-between the ground-truth grid points.

Generally, the error measures are lower for the piano pieces than for the operas. This is due to the design of our alignment method, which has been optimized for music with strong onsets that are present in piano music, but not in operas. For F08 and F09, the relative difference between the two triple errors ϵ_T and $\epsilon_{T,F}$ is higher than for the other pieces. This can partly be explained by the density of ground-truth annotations. The measure ground-truth grid of the opera pieces is much coarser compared to the beat annotations of the piano pieces, see for example F08 (0.37 #(GT)/s) compared to M24-2 (2.9 #(GT)/s) in Table 1. Furthermore, F08 exhibits the highest error. The piece is also the one with slowest tempo (having also less note onsets), and is partly performed as a recitative. This gives a high degree of freedom to the singer in shaping the local tempo of the performance, making it particularly difficult to achieve accurate alignments.

3.4. Identification of Problematic Versions

To identify if a specific recording introduces higher errors in the pairwise alignments, we included another performance by Biret (1990) into the set of recording for M17-4. This way, four triplets can be formed, each excluding one of the performances. Figure 5 shows the triple errors and the alignment errors for each of the four triplets. The experiment reveals that the triplet excluding the recording by Cohen (RubBirEza in Figure 5) leads to a lower difference between the triple error ϵ_T and the pairwise alignment error ϵ_P. Also worth noting is, that both triple errors and the pairwise alignment errors have the minimum for the same triplet. By using this strategy with a larger set of recordings, one could use the frame-wise triple error to identify problematic versions that generally lead to higher alignment errors in their pairwise alignments.

3.5. Comparing different alignment methods

Let $\mathcal{A}_1$ be our previously used alignment procedure. Furthermore, let $\mathcal{A}_2$ denote an alignment procedure only using chroma features without onset information. Figure 4a+b shows the mean error measures for $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively. Overall, by comparing Figure 4a and Figure 4b, we can see that using onset information in the features generally improves the alignments. This is reflected by most of the triple and the pairwise alignment errors. However, for the piece F08, although the pairwise alignment error ϵ_P is smaller for $\mathcal{A}_1$, the frame-wise triple error $\epsilon_{T,F}$ for $\mathcal{A}_1$ is larger than for $\mathcal{A}_2$. For

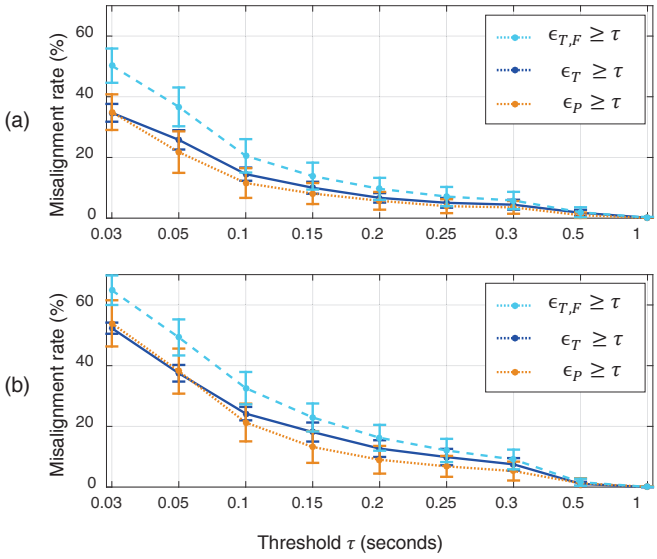

Fig. 6. Misalignment rates on frame-wise triple error $\epsilon_{T,F}$, triple error ϵ_T, and the pairwise alignment error ϵ_P on M17-4. Error bars indicate the standard deviation. **(a)** Alignment method $\mathcal{A}_1$ using chroma with onset information. **(b)** Alignment method $\mathcal{A}_2$ using chroma without onset information.

the piano pieces, all error measures are considerably smaller for $\mathcal{A}_1$ compared to $\mathcal{A}_2$, which shows that onsets are an important aspect for aligning piano music. Figure 6 shows the *misalignment rates* for M17-4 for the two alignment methods. It is defined as the percentage of time points in an alignment that have an error above a given threshold τ and is also commonly used to evaluate music alignment methods [3, 9]. The misalignment rate not only shows that $\mathcal{A}_1$ (Figure 6a) is more accurate than $\mathcal{A}_2$ (Figure 6b), but also reveals that the improvements are all below 0.3 seconds. Note that this conclusion can also be drawn by only considering the frame-wise triple error.

4. CONCLUSIONS

In this work, we introduced the triple error, a tool analyzing music alignment methods without the need of ground-truth annotations. It can be used when at least a triplet of recordings of the same piece of music is available. Although the triple error is only a necessary condition to indicate alignment errors, our experiments show similar tendencies as for the pairwise alignment error based on ground-truth annotations. We demonstrated that the triple error computed on a much finer frame grid can sometimes even indicate problematic positions in alignments that cannot be captured by ground-truth annotations. Also, we have shown that the triple error can be used to identify versions that are problematic in the pairwise alignments. Finally, despite its theoretical limitations, we indicated that the triple error can be used to compare different alignment methods.

In future work, we will apply the triple error on a large-scale analysis of datasets containing many versions of the same piece of music. We especially want to use it to identify problematic versions that cause the alignment method to fail. We will also investigate if it can be used to detect errors in ground-truth annotations. Finally, we want analyze whether alignment methods can be optimized or combined by using the triple error.

5. REFERENCES

[1] Roger B. Dannenberg and Christopher Raphael, "Music score alignment and computer accompaniment," *Communications of the ACM, Special Issue: Music Information Retrieval*, vol. 49, no. 8, pp. 38–43, 2006.

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

[3] Andreas Arzt and Gerhard Widmer, "Real-time music tracking using multiple performances as a reference," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, 2015.

[4] Siying Wang, Sebastian Ewert, and Simon Dixon, "Robust joint alignment of multiple versions of a piece of music," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, 2014, pp. 83–88.

[5] Hannah Ilea Robertson, *Testing a new tool for alignment of musical recordings*, Master's thesis, McGill University, 2013.

[6] Andreas Arzt, Harald Frostel, Thassilo Gadermaier, Martin Gasser, Maarten Grachten, and Gerhard Widmer, "Artificial intelligence in the Concertgebouw," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, Buenos Aires, Argentina, 2015, pp. 2424–2430.

[7] Craig Stuart Sapp, "Hybrid numeric/rank similarity metrics," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Philadelphia, USA, 2008, pp. 501–506.

[8] Robert J. Turetsky and Daniel P.W. Ellis, "Ground-truth transcriptions of real music from force-aligned MIDI syntheses," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Baltimore, Maryland, USA, 2003, pp. 135–141.

[9] Simon Dixon and Gerhard Widmer, "MATCH: A music alignment tool chest," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, London, UK, 2005.

[10] Arshia Cont, "A coupled duration-focused architecture for real-time music-to-score alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 974–987, 2010.

[11] Zhiyao Duan and Bryan Pardo, "A state space model for online polyphonic audio-score alignment," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 197–200.

[12] Roger B. Dannenberg and Ning Hu, "Polyphonic audio matching for score following and intelligent audio editors," in *Proceedings of the International Computer Music Conference (ICMC)*, San Francisco, USA, 2003, pp. 27–34.

[13] Meinard Müller, *Fundamentals of Music Processing*, Springer Verlag, 2015.

[14] 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, Taiwan, 2009, pp. 1869–1872.

[15] Sebastian Ewert and Meinard Müller, "Refinement strategies for music synchronization," in *Proceedings of the International Symposium on Computer Music Modeling and Retrieval (CMMR)*, Copenhagen, Denmark, 2008, vol. 5493 of *Lecture Notes in Computer Science*, pp. 147–165.