ANALYZING MEASURE ANNOTATIONS FOR WESTERN CLASSICAL MUSIC RECORDINGS

Christof Weiß[1] **Vlora Arifi-Müller**[1] **Thomas Prätzlich**[1]
Rainer Kleinertz[2] **Meinard Müller**[1]
[1] International Audio Laboratories Erlangen, Germany
[2] Institut für Musikwissenschaft, Saarland University, Germany
{christof.weiss, meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

This paper approaches the problem of annotating measure positions in Western classical music recordings. Such annotations can be useful for navigation, segmentation, and cross-version analysis of music in different types of representations. In a case study based on Wagner's opera "Die Walküre", we analyze two types of annotations. First, we report on an experiment where several human listeners generated annotations in a manual fashion. Second, we examine computer-generated annotations which were obtained by using score-to-audio alignment techniques. As one main contribution of this paper, we discuss the inconsistencies of the different annotations and study possible musical reasons for deviations. As another contribution, we propose a kernel-based method for automatically estimating confidences of the computed annotations which may serve as a first step towards improving the quality of this automatic method.

1. INTRODUCTION

Archives of Western classical music often comprise documents of various types and formats including text, symbolic data, audio, image, and video. Dealing with an opera, for example, one may have different versions of musical scores, libretti, and audio recordings. When exploring and analyzing the various kinds of information sources, the establishment of semantic relationships across the different music representations becomes an important issue. For a recorded performance, time positions are typically indicated in terms of *physical units* such as seconds. On the other hand, the musical score typically specifies time positions using *musical units* such as measures. Knowing the measure positions in a given music recording not only simplifies access and navigation [15, 19] but also allows for transferring annotations from the sheet music to the audio domain (and vice versa) [16]. Furthermore, a measure-based alignment of several performances enables cross-performance analysis tasks [12, 13].

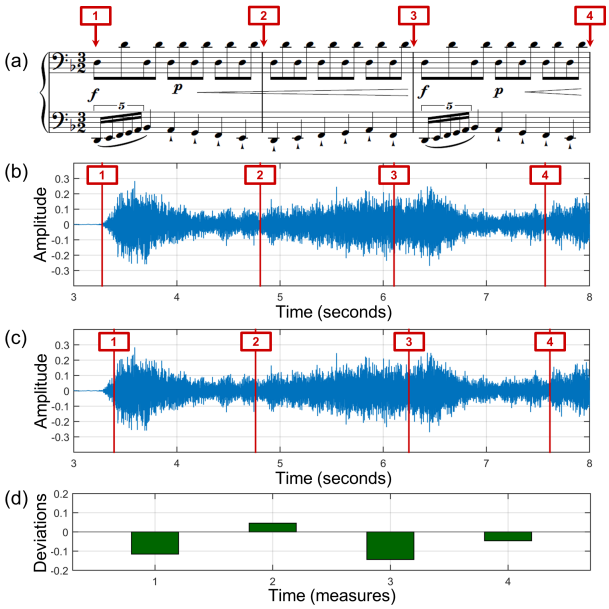

Figure 1. Two measure annotations for a Karajan performance of Wagner's opera "Die Walküre", first act, measures 1–3. **(a)** Piano reduction of the score. **(b)** Annotation A1. **(c)** Annotation A2. **(d)** Deviation of A1 from A2.

In this paper, we report on a case study based on the opera cycle "Der Ring des Nibelungen" WWV 86 by Richard Wagner where we consider the first act of the second opera "Die Walküre" (The Valkyrie). For this challenging scenario, we examine different types of measure annotations—either supplied by human annotators (*manual annotations*) or generated automatically using synchronization techniques (*computed annotations*). Figure 1 illustrates this scenario. Surprisingly, even the manual annotations (not to speak of the annotations obtained by automated methods) often deviate significantly from each other. As one contribution, we analyze such inconsistencies and discuss their implications for subsequent music analysis tasks. After describing the dataset and the annotation process (Section 2), we first analyze the properties of manual annotations stemming from different human annotators (Section 3). Subsequently, we evaluate computer-generated annotations that are derived from score-to-audio

© Christof Weiß, Vlora Arifi-Müller, Thomas Prätzlich, Rainer Kleinertz, Meinard Müller. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Christof Weiß, Vlora Arifi-Müller, Thomas Prätzlich, Rainer Kleinertz, Meinard Müller. "Analyzing Measure Annotations for Western Classical Music Recordings", 17th International Society for Music Information Retrieval Conference, 2016.

synchronization results (Section 4). Hereby, we examine correlations between inter-human inconsistencies and errors of the automated approach and identify musical reasons for the deviations. Finally, we propose a method to derive confidence values for the computed annotations from the synchronization procedure (Section 5).

2. DATA AND ANNOTATIONS

2.1 Music Scenario

Wagner's four-opera cycle "Der Ring des Nibelungen" WWV 86 is an exceptionally long work of about 14–15 hours duration. Because of its large scale and complex structure, it constitutes a challenging scenario for computer-assisted analysis methods [16]. In this paper, we consider the first act of "Die Walküre" WWV 86 B for a first case study. For analyzing such music, several issues are relevant. Work-related aspects such as motifs, instrumentation, chords, or coarse-scale harmonic structures as well as performance-related phenomena such as tempo, timbre, or loudness play a role. Furthermore, the relation between such aspects and the libretto may be of interest.

For their analyses, musicologists traditionally use the musical score which corresponds to the musical idea of the composer. Scores or piano reductions provide a compact overview of the musical content and are particularly suitable for harmony analysis. For performance-related aspects of the music, we need to analyze audio recordings. For this paper, we consider both types of data. Regarding symbolic data, we use a piano-reduced version of the score by Kleinmichel. [1] The sheet music is processed with OMR software (*Avid*™ *PhotoScore*) followed by manual correction using notation software. This piano-reduced score constitutes a kind of "harmonic excerpt" of the music. From the notation software, we export symbolic data types such as MIDI or MusicXML. For the audio domain, we consider an interpretation by Karajan with the Berlin Philharmonic (1966 Deutsche Grammophon, Berlin). [2] The duration of the first act in this recording is 67 minutes.

The types of music representations differ in the way how time and tempo are encoded. Audio recordings have a physical time axis usually given in seconds. In contrast, scores exhibit a musical time axis given in measures or beats. The physical length of a musical unit—such as a measure—depends on the *tempo* and the *time signature*. In operas, both tempo and time signature change frequently. To establish relations between the representations, we need to interconnect their time axes. One way to do this is to specify the measure positions in the audio recordings.

Such *measure annotations* may fulfill several purposes. First, they facilitate navigation and segmentation using musically meaningful units such as motifs, passages, or scenes [19]. Second, they enable the transfer of semantic annotations or analysis results from one domain to the other [16]. Third, *cross-version analysis* specifically uses the relation between different performances in order to stabilize analysis results [12].

2.2 Manual Annotations

To obtain measure annotations for our opera, we first consider a manual approach where five students with a strong practical experience in Western classical music annotated the measure positions for the full Karajan recording. We refer to these annotators as A1, . . . , A5. While following a vocal score [20] used as reference, the annotators listened to the recording and marked the measure positions using the public software *Sonic Visualizer* [2]. After finishing a certain passage, the annotators corrected erroneous or inaccurate measure positions. The length of these passages, the tolerance of errors, and the overall duration of the annotation process differed between the annotators. Roughly three hours were necessary to annotate one hour of music.

Beyond that, the annotators added comments to specify ambiguous measure positions. As musical reasons for such ambiguities, they mentioned tempo changes, fermatas, tied notes over barlines, or very fast passages. Furthermore, they reported performance-specific problems such as asynchronicities between orchestra and singers or masking of onsets through prominent other sounds. For some of these critical passages, the annotators reported problems arising from the use of a piano reduction instead of the full score.

Due to these (and other) difficulties, one can find significant deviations between the different annotations (see Figure 1 for an illustration). One goal of this paper is to analyze the annotation consistency and to uncover possible problems in the annotation process (Section 3).

2.3 Computed Annotations

The manual generation of measure annotations for music recordings is a time-consuming and tedious procedure. To automate this process, different strategies are possible. For example, one could start with a beat tracking algorithm and try to find the downbeats which yields the measure positions [17]. Moreover, beat information may help to obtain musically meaningful features [6]. For classical music, however, beat tracking is often not reliable [9, 10]. In [4], Degara *et al.* have automatically estimated the reliablity of a beat tracker.

In this paper, we follow another strategy based on synchronization techniques. The general goal of music synchronization (or audio-to-score alignment) is to establish an alignment between *musically corresponding time positions* in different representations of the same piece [1, 3, 5, 11]. Based on a symbolic score representation where measure positions are given explicitly, we use the computed alignment to transfer these positions to the audio recording.

[1] This piano reduction is publicly avaible on http://www.imslp.org.

[2] In our experiments, we also consider further performances yielding similar results as the ones reported for the Karajan performance.

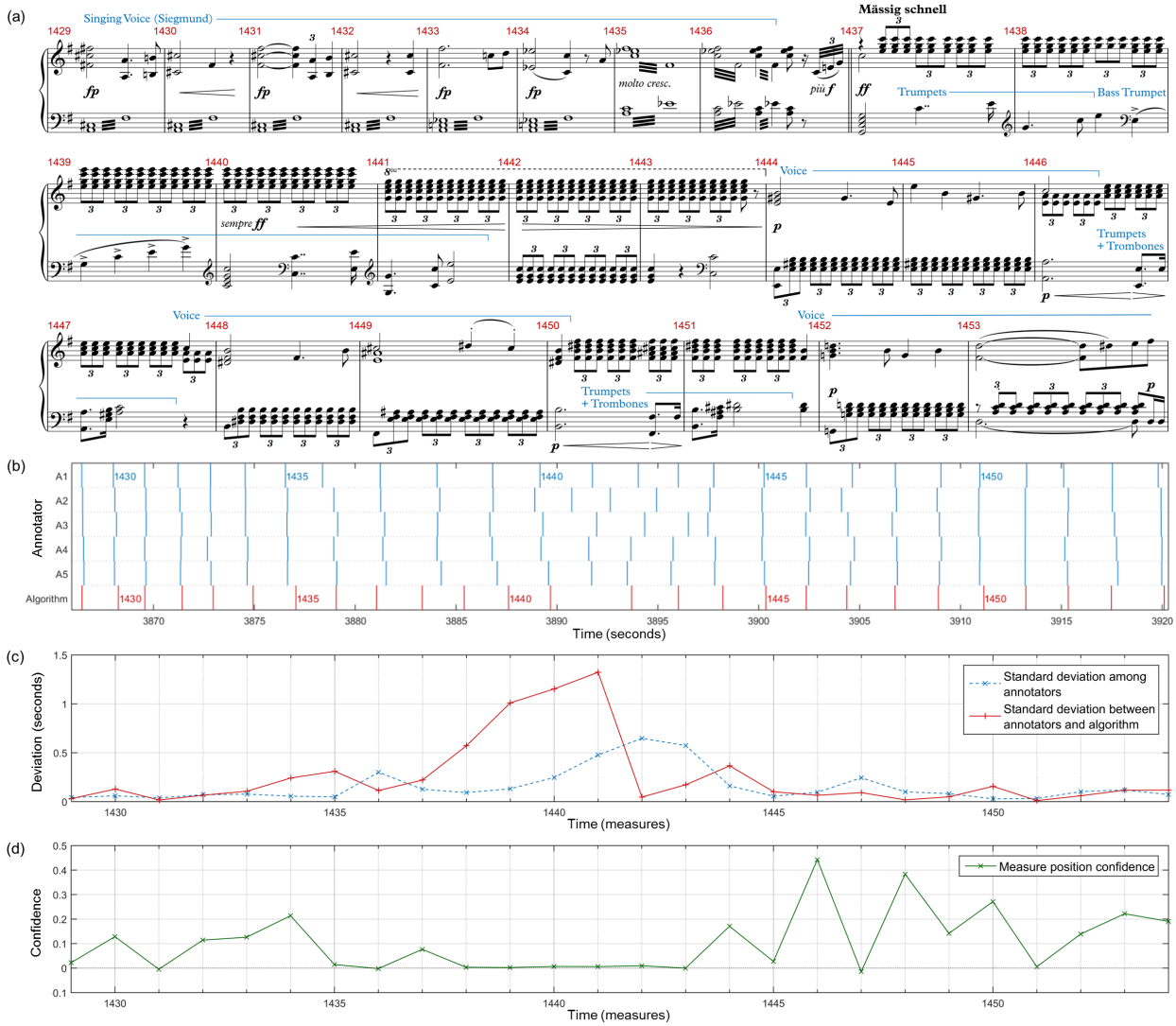

Figure 2. Measure annotations for a Karajan performance of R. Wagner's opera "Die Walküre", Act 1, Measures 1429–1453. **(a)** Piano reduction of the score (Kleinmichel). **(b)** Measure positions from manual (blue) and computed annotations (red). **(c)** Standard deviations among the manual annotations (blue dashed line) and between the mean manual annotation versus the algorithm's annotation (red solid line). **(d)** Measure position confidences derived from the similarity matrix.

In the following experiments, we use an alignment method based on Dynamic Time Warping (DTW) [15]. First, the audio recording and the symbolic score are transformed into a common feature representation. For this, we use CENS [15] features—a variant of chroma features which are well-suited for capturing the coarse harmonic progression of the music—combined with features capturing note onset information (see [7] for details). Using a suitable cost measure, DTW is applied to compute a cost-minimizing alignment between the two different versions [7]. As our opera recording is long (67 minutes), memory requirements and run time become an issue. To this end, we use a memory-efficient multiscale variant of DTW that allows for explicitly controlling the memory requirements [18]. The main idea of this DTW variant is to use rectangular constraint regions on which local alignments are computed independently using DTW. Macrae and Dixon [14] have used a similiar approach.

3. ANALYSIS OF MANUAL ANNOTATIONS

In our analysis, we first consider the manual annotations (see Section 2.2). As an example, Figure 2 shows a passage of "Die Walküre", Act 1. In Figure 2b, we plot the physical time position of the measures' beginning (horizontal axis) for the different annotators (vertical axis). At the beginning of this example, the annotators more or less agree. Sometimes, a single annotator slightly deviates from the others. As an example, annotator A1 sets an early position for measure 1436 compared to A2, . . . , A5. To quantify the overall disagreement for a specific measure, we calculate the standard deviation over the physical time position by all annotators. The blue dashed curve in Figure 2c shows this quantity for our exemplary passage. For example, one can see a small increase in measure 1436.

From measure 1440 on, the standard deviation consider-

ably increases over several measures. Looking at the annotations, we see that this disagreement does not stem from a single annotator but results from a substantial disagreement between all annotators. Annotator A2 exhibits the largest deviations by specifying the positions for the measures 1441–1443 much earlier than the other annotators. The confusion ends at measure 1445 for which the annotators seem to agree almost perfectly.

Looking at the score, we find possible hints for these deviations. For both measures 1436 and 1438–1443, the chord, voicing and instrumentation do not change with respect to the previous measure. In the measures 1437–1441, however, a prominent trumpet melody is present which probably served as an orientation for the annotators. In accordance with this assumption, we find the highest disagreement for the measures 1442 and 1443 where this melody has ended and only a constant and dense instrumentation of the C major chord is played. [3] Nevertheless, this observation does not explain the high deviations in measures 1440–1441. Listening to the recording, we noticed that the bass trumpet melody (left hand in m. 1439–1441) is covered by the orchestra and thus, practically not audible in this recording. Three of the annotators marked this problem as "masking". In measure 1444, a remarkable chord change (C major to E major) and a new melody in the tenor yield an orientation point where all annotators agree again.

By means of this examplary passage, we have already shown two important cues that may help humans to find measure boundaries: (1) Distinct harmonic changes and (2) salient melodic lines that can be followed easily. As for the second class, singing melodies seem to be more helpful than instrumental lines in the orchestra which are often superimposed by other instruments. For measures 1441 and 1445, we similarly find the present chord and instrumentation continued together with a melodic line. In the case of the trumpet line (m. 1441), the agreement is low. In contrast, the tenor melody (m. 1445) leads to high agreement. On the one hand, percussive speech components such as consonants and fricatives yield good temporal cues. On the other hand, solo voices play an important role in operas and often stand out musically as well as acoustically.

We have seen that humans may disagree substantially in their measure annotations. We now want to quantify such deviations on the basis of the full opera act. Since we do not have a "correct" ground truth annotation, we calculate for each measure the mean position across all manual annotations. Then, we calculate the offset of each annotation with respect to this mean position and plot a histogram over these offset values for all measures of the act. Figure 3 shows the resulting distributions for all five human annotators. In these plots, we observe typical offsets of about 0.1 seconds in both directions (measures annotated too early and too late). Deviations larger than 0.2 seconds are rare. Beyond this, we notice some systematic offsets towards one direction. For example, the distribution of A1 has its

[3] The full score shows 8th triplets (winds) and 16th arpeggios (strings). Our reduction focuses on the triplets that are hard to perceive in the audio.

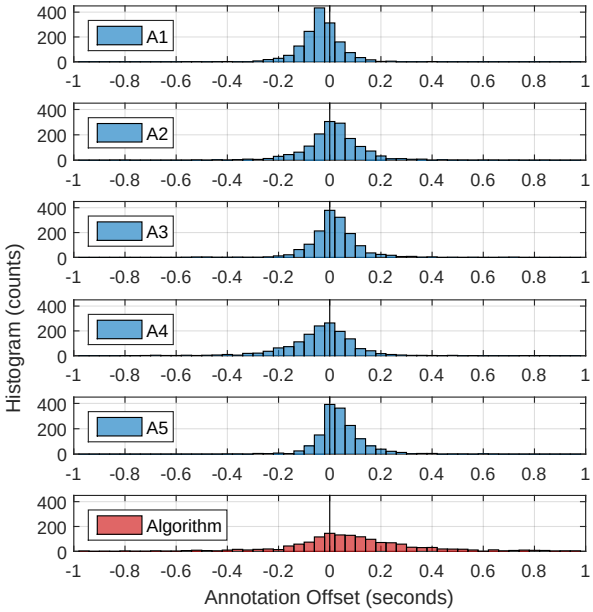

Figure 3. Histograms of annotation offsets for the individual annotations (full act). The deviations refer to the mean position of all annotations for the respective measure (labelled as zero). Positive offset values indicate "too late", negative values indicate "too early" positions compared to the mean position. The lowest plot refers to the computed annotation generated by our synchronization algorithm.

maximal bin at -0.04 seconds. For annotators A2, A3, and A5, the maximal bin is centered at zero but positive deviations are more frequent than negative ones. Overall, systematic offsets seem to be rather small. For single measures, deviations up to 0.2 seconds occur in both directions. In the following, we use the mean positions of all annotators as reference for evaluating our automatic approach.

4. ANALYSIS OF COMPUTED ANNOTATIONS

In this section, we analyze the computed annotations with respect to the manual annotations. Let us consider Figure 2 again. In the lowest row of Figure 2b (red), we show the measure positions as generated by the algorithm. The red curve in Figure 2c quantifies the deviation between the algorithm's and the average manual measure position. For the first measures 1429–1435, the computed positions seem to coincide with the manual annotations. Similarly, the annotations for the final measures 1445–1454 more or less agree. For the middle section, we find a different situation. In measures 1436–1441, the algorithm strongly deviates from the human annotators. For example, the position of measure 1441 is close to the human's position of measure 1440—a deviation of more than two seconds. Interestingly, the algorithm then produces a very long measure 1441 leading to a good coincidence with the manual annotations in measure 1442 again.

Looking at the score, we may find an explanation for this behaviour. The measures 1437–1443 (where the algorithm

strongly deviates) are harmonically restricted to a single C major triad. The listeners may have used the trumpet melody as cue to follow the rhythm. However, the trumpet only plays pitches from the C major chord which is present in the accompaniment. For chroma features which are the basis of the synchronization approach, these pitches contribute to the same chroma entries. For this reason, the chroma-based feature representation does not yield suitable cues for the matching process. For measures with clear harmonic change such as 1446 or 1448, we mostly see high agreement. Interestingly, one finds a small deviation for measure 1442 which is the most ambiguous measure among the human annotators. Here, we have to carefully interpret the figure since we use the mean manual annotations as reference. For measures with strong human disagreement, this mean position is not a reliable reference and, thus, the small deviation may be rather accidental.

In contrast, the relatively large deviation for measure 1444 seems surprising since we have a prominent harmonic change here (C major to E major chord). The situation becomes clearer when we listen to the audio recording. Actually, the onset of the singing voice (note B4) in measure 1444 is too early in this interpretation (by almost a quarter note) with respect to the chord change of the orchestra. The human annotators consistently followed the voice whereas the chroma-based synchronization method relies on the harmonic content dominated by the orchestra.

Let us consider Figure 3 again. The lowest plot (red) shows a histogram over the annotation offsets of the synchronization algorithm with respect to the mean manual annotation. This distribution is much flatter than the human ones. Large deviations such as the one in measure 1439 are more frequent for the automated approach. Furthermore, there is a remarkable systematic offset towards late measure positions. The majority of the annotations lies within a window of ± 0.3 seconds around the humans' mean position. For passages with strong disagreement, the algorithm finds back after a few measures—as for our example in Figure 2. Overall, we conclude that the automated approach does not reach the reliablity of the manual annotations but yields reasonable results for most measure positions.

5. CONFIDENCES FOR COMPUTED ANNOTATIONS

As we have seen from our previous analysis, there is a need for improving automated procedures. As a first step towards an improvement, we now introduce an approach for generating confidence values for the computed measure positions. Recall that the core of our method is a synchronization algorithm (see Section 2.3). Our idea is to use the local reliablity of the synchronization for estimating the confidence of the computed measure positions. Since the synchronization is based on chroma features, the change in harmony influences the quality of the alignment. Having similar chords in neighbouring measures usually results in similar chroma vectors. This often leads to situations

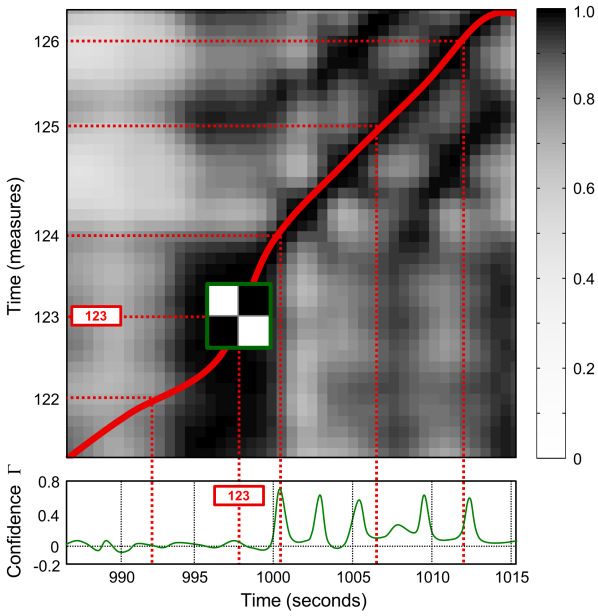

Figure 4. Estimation of measure position confidences (schematically). In the similarity matrix, we shift a checkerboard kernel along the warping path and calculate a confidence value for each measure position.

where the measure position is ambiguous. In contrast, measure boundaries that coincide with harmonic changes often lead to reliable measure annotations.

On the basis of this observation, we propose a novelty-based confidence measure. To compute the novelty score, we transfer our music recording to a sequence of chroma features $X = (x_1, \ldots, x_N)$ with a resolution of 10 Hz. Similarly, we compute a feature representation $Y = (y_1, \ldots, y_M)$ of the symbolic data (score). Then, we derive a similarity matrix $\mathcal{S} \in \mathbb{R}^{N \times M}$ from the two feature sequences using a cosine measure to compare feature vectors. The automated synchronization procedure (see Section 2.3) yields an alignment in terms of a warping path which we project on the given feature resolution. To estimate local confidence values for this warping path, we adapt an idea by Foote [8] who computes a novelty function by shifting a checkerboard kernel $\mathcal{K} \in \mathbb{R}^{K \times K}$ along the diagonal of a self-similarity matrix (SSM) and locally measures the similarity between $\mathcal{K}$ and the underlying region of the SSM. Here, we compute a novelty function by shifting the kernel along the *warping path* and locally measuring the similarity between $\mathcal{K}$ and the region of our similarity matrix $\mathcal{S}$. In our experiments, we use a kernel $\mathcal{K}$ of size $K = 10$ features (one second of the recording).

With this procedure and a subsequent normalization step, we obtain a curve $\Gamma : \{1, 2, \ldots, N\} \rightarrow [-1, 1]$ which measures the novelty of the local chroma vectors along the warping path. For a feature index $n \in \{1, \ldots, N\}$, a value of $\Gamma(n) \approx 1$ indicates high similarity between the local region of $\mathcal{S}$ and the structure of $\mathcal{K}$. Intuitively, we then expect a structural change in the features' properties. Musically spoken, $\Gamma(n) \approx 1$ implies clear change in local

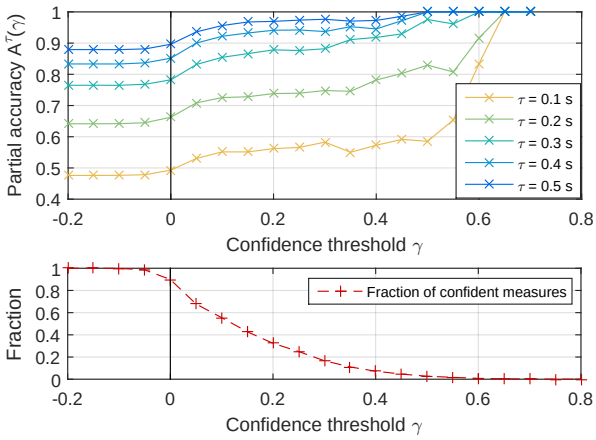

Figure 5. Confidence-dependent accuracy values for the entire act (1523 measures). For fixed tolerance values τ, the upper plot shows the partial accuracies $A^\tau(\gamma)$ of the computed measure annotations over the confidence threshold γ. The lower plot shows the fraction of measures under consideration.

harmony. At such positions, we expect the synchronization algorithm to work accurately. For $\Gamma(n) \approx 0$, there is little change in the features, which points to a harmonically homogenous situation in a neighbourhood of n. Finally, we evaluate Γ on those time instances that correspond to the measure positions. Figure 4 outlines this principle.

For quantitatively evaluating the computed annotations, we consider the mean of all five manual annotations as a reference. Using a tolerance parameter $\tau \in \mathbb{R}$, we regard a computed position to be correct if it lies within an interval of length 2τ centered at the reference position. Let $\mathcal{M} := \{1, 2, \ldots, L\}$ be the set of all measures and A^τ the fraction of correctly annotated measures with respect to τ. For $\tau = 0.2$ s, for example, a fraction $A^\tau = 62.5\,\%$ of the measures in $\mathcal{M}$ lies within this interval. We further define a subset $\mathcal{M}_\gamma := \{m \in \mathcal{M} \,|\, \Gamma(m) \geq \gamma\}$ which only includes measures with a confidence above the threshold $\gamma \in \mathbb{R}$. Additionally, we define a partial accuracy $A^\tau(\gamma)$ which only refers to the measures in $\mathcal{M}_\gamma$.

Figure 5 shows the results for the full first act of "Die Walküre". In the upper part, we show the curve $\gamma \to A^\tau(\gamma)$ for fixed τ. The lower plot displays the curve $\gamma \to |\mathcal{M}_\gamma|/|\mathcal{M}|$. In general, the accuracy increases for larger confidence thresholds γ. For $\tau = 0.3$ s (cyan curve), for example, $A^\tau(\gamma)$ improves from $78.2\,\%$ (for $\gamma = 0$) to $87.8\,\%$ (for $\gamma = 0.2$). At the same time, the fraction of considered measures decreases. To obtain accuracies $A^\tau(\gamma) > 90\,\%$, we end up evaluating less than $10\,\%$ of the measures (for $\tau = 0.3$ s). A good tradeoff seems to be at $\gamma = 0.1$ where we get up to $10\,\%$ increase of $A^\tau(\gamma)$ while still having half of the measure annotations included. These more "confident" measure positions may serve as a kind of anchor points for improving the quality of the automated approach. For example, one could replace the measure position with low consistency using linear interpolation or a smoothed tempo curve.

Finally, let us consider our running example again. Figure 2d shows the confidence values for this passage. We see that for the measures 1437–1443, the confidences are low due to the harmonic homongeneity (C major chord over seven measures). In contrast, we find high confidences for distinct chord changes as in measure 1444 (C major → E major), measure 1446 (E major → A minor), or measure 1448 (A minor → B major). Let us compare these values to the corresponding annotation consistency (red line in Figure 2c). For some of the high-confident measures (1446, 1448–1450, 1452–1454), the measure position is consistent with manual annotations. The situation is different for measure 1444. Here, our confidence value is high but the position deviates from the manual annotations. Remembering the audio properties discussed in Section 4, we can understand this behaviour. Here, the human annotators consistently follow the entry of the voice which is too early compared to the orchestra's onset. Thus, the high confidence indicates a good measure position which is correct with respect to *harmony*. The deviation from the manual annotations arises from the asynchronicity.

In general, we can only draw conclusions for measures with high confidence $\Gamma(m)$. A low confidence does not necessarily indicate a bad estimate of the measure position. In Figure 2, measures 1431 and 1445 are examples for such a behaviour, where we find low Γ-values but high consistency of measure positions with manual annotations.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we analyzed different types of measure annotations for an opera recording. For manual annotations generated by different human annotators, we identified musical challenges which can lead to inconsistencies among the annotators. In contrast, harmonic changes and melodic lines seem to be important cues for the listeners to accurately locate measure boundaries. Furthermore, we analyzed measure annotations generated by a computer using score-to-audio alignment. This approach provides useful results but is less accurate than manual annotations. In particular, harmonic homogeneity can be problematic for chroma-based approaches. Based on this observation, we automatically estimate the confidence of the computed annotations. To this end, we shift a checkerboard kernel along the warping path. The resulting confidence values seem to be useful for identifying reliable measure position. Thus, they may serve as a first step towards improving synchronization-based annotation strategies.

7. ACKNOWLEDGMENTS

We thank all students involved in the annotation work. This work was supported by the German Research Foundation (DFG MU 2686/7-1, DFG KL 864/4-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg FAU and Fraunhofer Institute for Integrated Circuits IIS.

8. REFERENCES

[1] Andreas Arzt and Gerhard Widmer. Real-time music tracking using multiple performances as a reference. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 357–363, Málaga, Spain, 2015.

[2] Chris Cannam, Christian Landone, Mark Sandler, and Juan Pablo Bello. The Sonic Visualiser: A visualisation platform for semantic descriptors from musical signals. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 324–327, Victoria, Canada, 2006.

[3] 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)*, pages 27–34, San Francisco, USA, 2003.

[4] Norberto Degara, Enrique Argones Rúa, Antonio Pena, Soledad Torres-Guijarro, Matthew E. P. Davies, and Mark D. Plumbley. Reliability-informed beat tracking of musical signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):290–301, 2012.

[5] 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.

[6] Daniel P.W. Ellis, Courtenay V. Cotton, and Michael I. Mandel. Cross-correlation of beat-synchronous representations for music similarity. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 57–60, Las Vegas, USA, 2008.

[7] 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)*, pages 1869–1872, Taipei, Taiwan, 2009.

[8] Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 452–455, New York, USA, 2000.

[9] Peter Grosche, Meinard Müller, and Craig Stuart Sapp. What makes beat tracking difficult? A case study on Chopin Mazurkas. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 649–654, Utrecht, The Netherlands, 2010.

[10] André Holzapfel, Matthew E.P. Davies, José R. Zapata, João Oliveira, and Fabien Gouyon. On the automatic identification of difficult examples for beat tracking: towards building new evaluation datasets. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 89–92, Kyoto, Japan, 2012.

[11] 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*, 19(8):2385–2397, 2011.

[12] Verena Konz, Meinard Müller, and Rainer Kleinertz. A cross-version chord labelling approach for exploring harmonic structures – a case study on Beethoven's Appassionata. *Journal of New Music Research*, pages 1–17, 2013.

[13] Cynthia C.S. Liem and Alan Hanjalic. Expressive timing from cross-performance and audio-based alignment patterns: An extended case study. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 519–524, Miami, USA, 2011.

[14] Robert Macrae and Simon Dixon. Accurate real-time windowed time warping. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 423–428, Utrecht, The Netherlands, 2010.

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

[16] Kevin R. Page, Terhi Nurmikko-Fuller, Carolin Rindfleisch, David M. Weigl, Richard Lewis, Laurence Dreyfus, and David De Roure. A toolkit for live annotation of opera performance: Experiences capturing Wagner's ring cycle. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 211–217, Málaga, Spain, 2015.

[17] Hélène Papadopoulos and Geoffroy Peeters. Joint estimation of chords and downbeats from an audio signal. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):138–152, 2011.

[18] Thomas Prätzlich, Jonathan Driedger, and Meinard Müller. Memory-restricted multiscale dynamic time warping. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 2016.

[19] Thomas Prätzlich and Meinard Müller. Freischütz Digital: a case study for reference-based audio segmentation of operas. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 589–594, Curitiba, Brazil, 2013.

[20] Richard Wagner. *Die Walküre. Vocal score based on the complete edition*. Schott Music (ed. Egon Voss), Mainz, Germany, 2013.