ROBUST AND ACCURATE AUDIO SYNCHRONIZATION USING RAW FEATURES FROM TRANSCRIPTION MODELS

Johannes Zeitler, Ben Maman and Meinard Müller
International Audio Laboratories Erlangen, Germany
{johannes.zeitler, ben.maman, meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

In music information retrieval (MIR), precise synchronization of musical events is crucial for tasks like aligning symbolic information with music recordings or transferring annotations between audio versions. To achieve high temporal accuracy, synchronization approaches integrate onset-related information extracted from music recordings using either traditional signal processing techniques or exploiting symbolic representations obtained by data-driven automated music transcription (AMT) approaches. In line with this research direction, our paper introduces a high-resolution synchronization approach that combines recent AMT techniques with traditional synchronization methods. Rather than relying on the final symbolic AMT results, we show how to exploit raw onset and frame predictions obtained as intermediate outcomes from a state-of-the-art AMT approach. Through extensive evaluations conducted on piano recordings under varied acoustic conditions across different transcription models, audio features, and dynamic time warping variants, we illustrate the advantages of our proposed method in both audio–audio and audio–score synchronization tasks. Specifically, we emphasize the effectiveness of our approach in aligning historical piano recordings with poor audio quality. We underscore how additional fine-tuning steps of the transcription model on the target dataset enhance alignment robustness, even in challenging acoustic environments.

1. INTRODUCTION AND RELATED WORK

Aligning different versions of a musical piece is a common task in music information retrieval (MIR). For example, score–audio synchronization with the objective to align score-based note information with time positions of an audio recording is used in automatic score following [1, 2], score-informed audio decomposition techniques [3], or the derivation of note labels for the training and evaluation of automated music transcription (AMT) systems [4]. Aligning different audio recordings of the same musical piece (audio–audio synchronization) enables applications like track switching [5], cross-version analysis [6], automated accompaniment of instrumentalists using existing backing tracks [7, 8], and the transfer of annotations from one recording to another [9, 10].

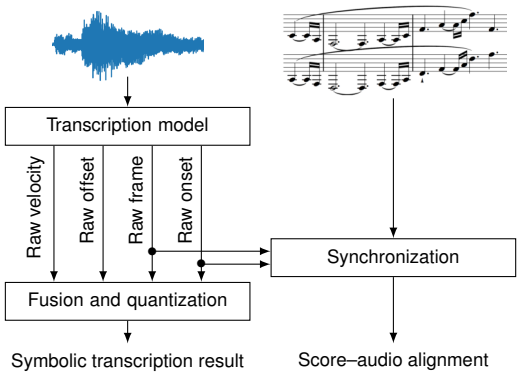

Figure 1: Schematic overview of the proposed audio–score synchronization pipeline using raw features from a transcription model.

Alignment pipelines based on dynamic time warping (DTW) typically use chroma or onset features, or a combination of both [11, 12]. While such features can easily be obtained from symbolic score information, they need to be estimated from audio recordings. Traditionally, many alignment pipelines rely on features estimated with classical signal processing methods, e.g., using a constant-Q transform [13] or a multirate filterbank [11, 14]. With the advancements in deep learning (DL) techniques, several systems for multi-pitch estimation (MPE) [15–20] as well as learning-based methods for onset estimation [21–24] have been introduced. Along with the creation of large datasets of pairs of audio recordings and note labels such as MAESTRO [25] or MusicNet [16], modern transcription models precisely estimate note on- and offset, as well as velocity and pedaling information [26–28].

In this work, we investigate the advantages of using features estimated by AMT systems for audio–audio and audio–score alignment tasks. We demonstrate how to leverage intermediate predictions from transcription models for aligning audio recordings and symbolic representations, as illustrated in Figure 1. In particular, we investigate alignments within a carefully curated dataset of the first movements of the 32 piano sonatas by Ludwig van Beethoven, with all sonatas performed by eleven artists, encompassing live performances, historic recordings with low audio quality, performances on historic instruments,

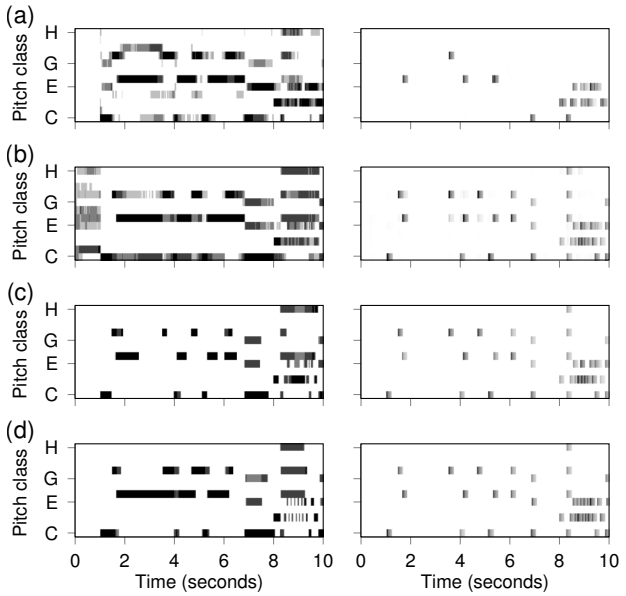

Figure 2: Chroma (left) and DLNCO (right) features for Beethoven's Appassionata (Op. 57-1) played by F. Dupree. (a) `FB_RAW`. (b) `T1_RAW`. (c) `T1_SYM`. (d) `DK_SYM`.

and modern studio-quality recordings. By analyzing the alignment precision and robustness across different feature extractors and representations, synchronization algorithms, and audio versions, we demonstrate that our approach allows for robust synchronization of real-world data.

The outline of this paper is as follows. First, in Section 2, we describe the synchronization pipeline and discuss its components, followed by an introduction to our dataset of Beethoven's piano sonatas in Section 3. Since there are no reference alignments available for the audio recordings in this real-world dataset, we rely on heuristics for evaluation, see Section 4. In Section 5, we experimentally show how using raw intermediate features from transcription models increases the alignment stability, as compared to using the final symbolic transcription results. Furthermore, we give detailed insights into the peculiarities of aligning datasets encompassing historic music recordings and demonstrate how to adapt to such data. We conclude in Section 6 with an outlook on future work.

2. SYNCHROZINATION PIPELINE

In this section, we provide an overview of the synchronization pipeline. First, we describe the feature extractors and types of feature representations, before we outline the DTW-based alignment step. An overview of all elements in the pipeline and their abbreviations is provided in Table 1. Following the notation in [26], we distinguish between two types of features: *frame* features encode when a note is active (in the piano case, this corresponds to the time until a key is released, or until the sustain phase ends), and *onset* features encode only the beginning of a note (in the piano case when a key is pressed).

Feature Extraction Model	
`FB`	Filterbank
`T1`	Onsets and frames transcription model [4]
`T2`	High-resolution transcription model [27]
`DK`	Disklavier
Feature Representation	
`RAW`	Continuous pitch and onset probabilities
`SYM`	Thresholded and discretized pitches and onsets
Alignment Technique	
`O`	Onset features using standard DTW
`OF`	Onset and frame features using MrMsDTW from [14]

Table 1: Overview of short notation for all components of the processing chain.

2.1 Feature Extraction Model

Filterbank. Before the advent of today's DL-based feature extraction models, traditional signal processing techniques were a common way to extract features from audio recordings. For example, the standard implementation of Sync Toolbox [14] uses a multirate filterbank (`FB`) to estimate frame-wise note activity and onsets.

Transcription Model. In recent years, AMT systems based on DL have shown significant improvements in performance [29]. One of the ground-breaking architectures is the *Onsets and Frames* architecture by Hawthorne et al. [26], which has separate prediction heads to estimate onsets and frames. Maman et al. [4] proposed a strategy to train a model based on the onsets and frames architecture on diverse and unaligned pairs of audio data and musical scores. This has led to improved performance on unseen datasets, generalizing across instrumentations, acoustic conditions, and styles. In the following, we refer to the transcription model from [4], trained on MusicNet [16] with re-aligned labels, as `T1`. Kong et al. [27] extend the onsets and frames architecture by additionally modeling sustain pedal activity, therefore providing more robust training in the presence of misaligned offset information. We refer to the transcription system from [27], trained on MAESTRO [25], as `T2`. [1]

Disklavier. Certain datasets such as MAESTRO [25] include pairs of audio recordings and reference note information by having the pieces performed on a Disklavier. We refer to features directly extracted from the symbolic Disklavier track as `DK`, and use them as an upper bound for the performance of an MPE feature extractor.

2.2 Feature Representation

Raw features. For each sequence of input audio, the feature extractors `FB`, `T1`, and `T2` predict continuous pitch- and frame-wise probabilities $\mathbf{P}_{\text{raw}}^{\text{frame}}, \mathbf{P}_{\text{raw}}^{\text{onset}} \in [0,1]^{88\times N}$ for frame activity and note onset, respectively. These feature matrices can be thought of as `RAW` features and are commonly stored in a pianoroll-like representation for 88 pitches and N time frames (`T1` and `T2` additionally predict

[1] Note that we do not include transcription models that directly output a tokenized sequence of MIDI messages (where we can not access raw pitch probabilities), such as the MT3 model by Hawthorne et al. [30].

such probabilities for note velocity and offset). Figure 2a/b illustrates RAW features for a piece from the ASAP dataset [31], computed by the FB and T1 feature extractors .

Symbolic features. In AMT, the raw predictions for frames, onsets, and offsets are fused and quantized in a postprocessing step, yielding binary estimates about which keys have been pressed. In particular, note sustain (frame activity) is conditioned on a previously occurring note onset [26,27]. This postprocessing step outputs a sequence of symbolic control messages for note onset and offset events, with additional control messages for pedal information in the case of T2. We denote these binary and symbolic-like features as SYM features and, for usage in our synchronization pipeline, store these in the form of two discretized pianorolls $\mathbf{P}_{\text{sym}}^{\text{frame}}, \mathbf{P}_{\text{sym}}^{\text{onset}} \in \{0,1\}^{88\times N}$ for frame activity and onset events, respectively. Note that in the case of Disklavier (DK), no RAW features are available and thus only SYM features are used. Figure 2c/d illustrates SYM features for T1 and DK.

Comparison. In Figure 2 we qualitatively compare RAW features from FB and T1 as well as SYM features from T1 to the DK reference. While FB_RAW in Figure 2a shows many false positive chroma events and misses many onsets compared to DK in Figure 2d, the chroma features of T1_RAW in Figure 2b are relatively stable and onsets perfectly coincide with DK. Thresholding the RAW transcription results to T1_SYM features (Figure 2c) yields varying and often shortened note durations in the chroma representation compared to DK, indicating possible instabilities when using these features for computing an alignment.

2.3 Alignment Technique

We use two variants of DTW to compute the optimal alignment between two feature sequences.

Onset features. As a first approach and in line with previous work [4, 32], we use only onset features and convert them to a twelve-dimensional pitch class representation. Using the Euclidean distance function, we compute the cost matrix between the onset feature sequences of the two versions to be aligned. We use standard DTW with unit steps in the horizontal, vertical, and diagonal direction with step weights (1.5, 1.5, 2) to compute the minimum cost path between the two sequences [12]. We refer to this approach, using only onset features, as O.

Onset and frame features. As a second alignment variant, we choose a high-resolution approach [11] that combines frame and onset features. Using frame features yields robustness on the coarse temporal level, while onset features provide precision on the fine level by precisely aligning note onsets [24]. In this approach, we again convert frame and onset features into pitch class representations and additionally add a decay to the onset features. We refer to [11] for a description of these decaying locally normalized chroma onset (DLNCO) features. Next, we compute separate cost matrices for frame features (using the cosine distance) and for onset features (using the Euclidean distance). Afterward, we add the two cost matrices for frame and onset features and use DTW with step weights (1.5, 1.5, 2) to compute the optimum alignment path on the combined cost matrix. We refer to [14, 33] for an efficient multi-resolution and multi-scale implementation of DTW. We denote the described approach, using a combination of onset and frame features, as OF. Note that we do not consider using only frame features (commonly called chroma features), as previous work has shown a lack of precision in this case. For example, Ewert et al. observe a 100 % increase of the alignment error when using frame features instead of combined frame and onset features for the case of piano music, where onsets are clearly defined [11].

ID	Performer	Year	Duration
AS35	Artur Schnabel	1935	03:33:35
FG58	Friedrich Gulda	1958	03:34:00
FJ62	Fritz Jank	1962	03:41:26
WK64	Wilhelm Kempff	1964	03:45:31
FG67	Friedrich Gulda	1967	03:25:02
VA81	Vladimir Ashkenazy	1981	03:46:27
DB84	Daniel Barenboim	1984	03:58:37
JJ90	Jeno Jando	1990	03:39:14
AB96	Alfred Brendel	1996	03:52:28
MB97	Malcolm Bilson et al.	1997	03:46:08
MC22	Muriel Chemin	2022	04:05:11
		Total	41:07:45

Table 2: Overview of audio versions in the BPSD. The versions with identifiers AS35, FG58, FJ62, and WK64 are in the public domain and are freely accessible within the BPSD. Durations given in hh:mm:ss.

3. DATASETS

In our experiments, we consider the case of piano music, as there are large-scale datasets available [25, 31, 34], note onsets are well-defined, reference note information can be obtained from performances on a Disklavier, and many transcription models are primarily trained on piano music [26, 27]. To this end, we evaluate alignment accuracy not only in acoustically controlled scenarios such as MAESTRO. Instead, we consider a much more challenging scenario using real-world piano recordings under complex acoustic conditions, which we find in a dataset of Beethoven's piano sonatas [35]. The 32 piano sonatas by Ludwig van Beethoven are recognized as pivotal works in Western classical music and hold a significant place in cultural history. Being one of the most performed and recorded corpus of classical music, alignments between a multitude of different versions can be studied.

3.1 Beethoven Piano Sonata Dataset

As a main evaluation corpus, we choose the Beethoven Piano Sonata Dataset (BPSD) [35], which comprises eleven complete audio recordings of the first movements of all 32 piano sonatas, along with sheet music in machine-readable format. An aspect of central importance is the coherent structure of the dataset: all audio versions and the symbolic sheet music share the same musical timeline, which was enforced by manually editing the score and audio versions. Thus, there is no incoherence due to, e.g., additional

or missing repetitions. The BPSD includes over 41 h of audio recorded under various acoustic conditions, being far more diverse than common piano datasets [25, 34]. For example, MAESTRO was entirely performed on Yamaha Disklaviers, and training on MAESTRO does not provide good generalization on other datasets [4, 32, 36]. In contrast, the BPSD comprises modern studio recordings in high audio quality, vintage recordings published on vinyl, including pitch drift due to wobbling of the vinyl records, performances on historical instruments such as the fortepiano, and significant deviations from today's standard tuning frequency of 440 Hz (A4). Measure positions were annotated manually for all 32 sonatas recorded by Wilhelm Kempff in 1964 (`WK64`). An overview of the eleven audio versions in the BPSD is provided in Table 2.

3.2 ASAP

To be able to use reference note information from Disklavier recordings in our experiments, we additionally leverage the ASAP dataset [31].To achieve consistency across all experiments, we identify the performances of the first movements of Beethoven's piano sonatas in ASAP which share the same structure as recordings in the BPSD. This subset consists of 13 individual recordings with a total length of 103 min.

4. QUANTIFYING SYNCHRONIZATION ACCURACY

In this section, we describe the heuristics used to assess the accuracy of our score–audio and audio–audio synchronization pipelines. We refer to [37] for a detailed discussion about the analysis of synchronization accuracy without ground-truth annotations.

4.1 Notation

We first introduce some notation for aligning time points between two different versions V_1 and V_2 of a piece. We assume that these versions have continuous time axes $[0, T_1]$ and $[0, T_2]$, which can either be in physical time (for audio recordings, in seconds) or in musical time (for score-related data, in measures). From the alignment algorithms described in Section 2.3, we obtain a monotonous mapping function $\mathcal{M}^{V_1 \to V_2} : [0, T_1] \to [0, T_2]$ to transfer time instants from the timeline of one version to the other. Note that even though the alignment result obtained from DTW maps discrete time axes, our assumption of having continuous time axes can be obtained by using suitable interpolation techniques, see [37].

4.2 BPSD: Measure Transfer

In the following, we consider three versions: $V_1 = S$ being a score, and $V_2 = A_1$ and $V_3 = A_2$ being different audio versions of the same piece. We choose A_2 to be the recordings by Wilhelm Kempff (`WK64`), for which we have access to manually annotated measure positions t_{A_2}. Using audio–audio synchronization, we obtain a mapping $\mathcal{M}^{A_2 \to A_1}$ to transfer these measure positions to the first audio version A_1. Similarly, we transfer measure positions t_S obtained from the score S to the first audio A_1 using a mapping $\mathcal{M}^{S \to A_1}$. In a last step, as illustrated in Figure 3a, we calculate the absolute error of the measure positions transferred from S and A_2:

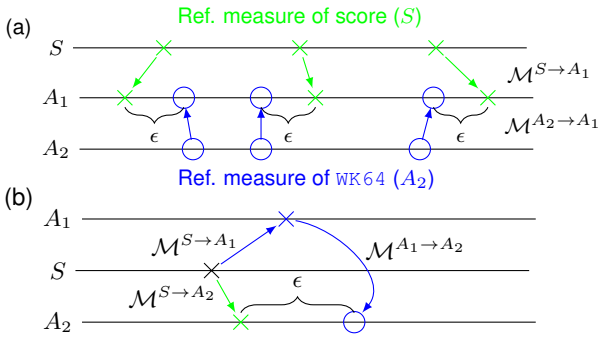

Figure 3: Schematic illustration of (a) measure transfer and (b) note onset transfer heuristics.

$$\epsilon = \left| \mathcal{M}^{S \to A_1}(t_S) - \mathcal{M}^{A_2 \to A_1}(t_{A_2}) \right| . \quad (1)$$

4.3 ASAP: Note Onset Transfer

In order to evaluate score–audio synchronization on the ASAP dataset, we can not resort to the heuristic described in Section 4.2, as there are no reliable manual measure annotations available. Therefore, we use an approach that transfers note onsets, illustrated in Figure 3b, to assess the synchronization accuracy.

First, we obtain audio features for two different audio recordings A_1, A_2 of the same piece, as well as features for the score S. For each version, we perform score–audio synchronization to obtain alignment functions $\mathcal{M}^{S \to A_1}$ and $\mathcal{M}^{S \to A_2}$ from musical to physical time. Using these mapping functions, we map note event onsets t_S in musical time from the score to the physical time of the audio recordings. In a second step, we transfer the aligned onset positions from the first to the second audio using audio–audio synchronization via the mapping function $\mathcal{M}^{A_1 \to A_2}$ and compute the absolute error

$$\epsilon = \left| \mathcal{M}^{S \to A_2}(t_S) - \mathcal{M}^{A_1 \to A_2}\left(\mathcal{M}^{S \to A_1}(t_S)\right) \right| \quad (2)$$

between these transferred time points and the ones obtained from score–audio synchronization.

In both heuristics, we assume that the synchronization accuracy is high if the time instances transferred via two different branches have small deviations. Note that this is only a *necessary* and not a *sufficient* condition for alignment quality; nevertheless, this metric gives a good indicator of the alignment performance (see also [37]).

5. EVALUATION

While our main focus is on the BPSD due to its realism and diversity, we first analyze synchronization accuracy on the ASAP dataset in order to compare features estimated from audio recordings to those derived from the Disklavier reference. In the next step, we evaluate the alignment performance on the BPSD across all audio versions. Finally, we

Feature	Mean	Median	Conf. 90	Conf. 95
T1_SYM_O	89	**0**	281	467
T1_SYM_OF	66	12	153	293
T1_RAW_OF	29	8	64	146
T2_SYM_O	31	**0**	102	192
T2_SYM_OF	22	2	49	111
T2_RAW_OF	**21**	7	**43**	**99**
DK_SYM_O	37	**0**	130	216
DK_SYM_OF	25	2	57	123
FB_RAW_OF	64	20	146	268

Table 3: ASAP: Absolute error in milliseconds for note onset transfer heuristic.

conduct a detailed analysis of the performance on individual versions, identify problematic recordings, and illustrate how to improve alignment robustness by adapting a transcription model to the target data.

5.1 ASAP: Estimated Features vs. Reference Notes

First, we evaluate synchronization accuracy on the ASAP dataset using our note onset transfer approach described in Section 4.3. For each pair of audio files, we calculate the mean, median, 90 and 95 percentiles of the absolute alignment error in ms, and report averaged results in Table 3.

Analyzing the median absolute error, which we consider an indicator for the achievable accuracy under average conditions, we find perfect alignments ($\epsilon = 0\,\mathrm{ms}$) for at least $50\,\%$ of all note onsets when using only symbolic onset features (SYM_O) for the transcription models T1 and T2, as well as the Disklavier DK. To assess the methods' robustness and the severeness of outliers, we next investigate the $95\,\%$ quantiles of the absolute alignment error. Using only symbolic onset features and standard DTW (SYM_O) yields the highest errors for all T1 (467 ms), T2 (192 ms), and DK (216 ms) variants, indicating a lack of robustness despite excellent median accuracy.

In the next step, we jointly use the frame and onset information from the symbolic features (SYM_OF) and observe a slight rise in the median error to 12 ms for T1, and to 2 ms for T2 and DK, respectively. While this indicates that the best achievable precision slightly deteriorates, we monitor a significant reduction of the $90\,\%$ and $95\,\%$ confidence intervals by approximately $50\,\%$ for all SYM_OF variants, indicating a vastly improved robustness towards outliers when combining frame and onset features in the computation of the alignment.

Lastly, we directly use the intermediate predictions for frames and onsets (RAW_OF) in our alignment pipeline. While the median absolute error is comparable to the one based on symbolic features, we again observe a significant decrease in the $90\,\%$ and $95\,\%$ confidence intervals. Using RAW_OF features in the T2 transcriber yields the lowest mean (21 ms) and confidence intervals (43 and 99 ms), even outperforming the usage of reference note information obtained from the Disklavier. [2] We illustrate this finding with the intuitive example of a chord where notes are not played simultaneously, either due to a playing mistake or as a stylistic element, leading to a deviation of symbolic and actually performed note order. While the DK features strictly assign each note onset to one particular time frame and thus cause alignment instabilities in the given example, the continuous RAW predictions can smoothly cover neighboring time frames and thus allow for a robust alignment.

Feature	Mean	Median	Conf. 90	Conf. 95
T1_SYM_O	66	17	115	234
T1_SYM_OF	52	20	102	160
T1_RAW_OF	**41**	**12**	**70**	**121**
T2_SYM_O	109	20	272	466
T2_SYM_OF	56	14	138	251
T2_RAW_OF	47	15	97	207
FB_RAW_OF	44	20	**70**	128

Table 4: BPSD: Absolute error in milliseconds for measure transfer heuristic.

5.2 BPSD: General Performance

Next, we analyze the overall matching of score–audio and audio–audio synchronization on the more realistic and more diverse BPSD by using the measure-transfer heuristic as described in Section 4.2. In Table 4, we again report the mean, median, 90 and $95\,\%$ confidence intervals for the absolute error between measure positions obtained from score–audio and audio–audio transfer.

Analyzing the median absolute error in Table 4, all features yield a precision between approximately 12 ms and 20 ms, without a clear tendency towards one particular method. However, it is the robustness (measured by the $90\,\%$ and $95\,\%$ confidence intervals) where we find a clear trend: using only onsets from symbolic features (SYM_O) yields large alignment outliers, with the $95\,\%$ confidence interval of the absolute error being 234 ms for T1 and even 466 ms for T2. Using additional frame features (SYM_OF) lowers the $95\,\%$ confidence interval to 160 ms and 251 ms, respectively. In line with our observations on the ASAP dataset, using intermediate transcription results (RAW_OF) further reduces the mean as well as the confidence intervals for both transcription models. The T1 transcriber, which was trained on audio from the acoustically diverse MusicNet [16] dataset, exhibits significantly lower errors than T2 (121 ms vs. 207 ms for the $95\,\%$ conf. interval), which was trained only on MAESTRO. While using filterbank features (FB) resulted in relatively high errors on ASAP, on the BPSD we observe metrics that are similar to those of the T1 transcriber, and considerably better than those of the T2 model. This indicates a lack of robustness of the DL-based transcription models on the diverse acoustic conditions of the BPSD, which we will investigate and mitigate in the following section.

[2] We note that the T2 model was trained on MAESTRO [25], which is the basis of ASAP [31]. Therefore, a separation for train and test data is not guaranteed for T2 in the experiments on ASAP. However, the DK features nevertheless are the upper limit of the achievable transcription accuracy.

5.3 BPSD: Detailed Analysis and Finetuning

To further investigate why the alignment pipelines using transcription features (T1,T2) do not yield significantly better results on the BPSD than those features using the filterbank baseline (FB), we further break down the investigation to the BPSD's individual audio versions. Transcription models (and DL systems in general) are known to exhibit a degraded performance when there is a domain shift between the test data and the training data [32, 36]. Such effects can be caused by poor audio quality in general, or, for music data, by a difference in timbre or tuning.

Identifying problematic versions. We restrict our analysis to raw frame and onset features (RAW_OF) from the filterbank (FB) and the T1 transcriber, as these showed the overall most robust results on the complete BPSD (see Section 5.2) and illustrate the median and 95 % confidence intervals for all audio versions of the BPSD in Table 5. Analyzing the 95 % confidence values for T1_RAW_OF in Table 5, we identify two problematic versions, namely the 1958 recordings by Friedrich Gulda (FG58) with 788 ms and the 1997 recordings by Malcolm Bilson et al. (MB97) with 146 ms. By inspection of the recorded pieces, we find two different reasons for the alignment instabilities.

Musical and acoustic reasons for instabilities. Friedrich Gulda recorded his first cycle of Beethoven's Piano Sonatas (FG58) over a relatively long time span between 1950 and 1958, playing different pianos in different environments. Among the FG58 recordings, we identify Sonata No. 26 ("Les adieux") and No. 29 ("Hammerklavier") as especially problematic, showing large differences in tuning, along with high background noise.

The pianist Malcolm Bilson (MB97) is committed to historically informed performance practice. His interpretations on historical instruments introduce a novel approach to performance in an era predominantly defined by the use of modern instruments. Malcolm Bilson and colleagues recorded their 1997 cycle of Beethoven's Piano Sonatas on nine fortepianos, including original historical instruments. Compared to modern pianos, the overall sound of fortepianos is significantly different, due to different mechanics, strings, and resonance bodies. Furthermore, the timbre varies across registers, e.g., bass notes sound fundamentally different compared to high-octave notes, and the reference pitch deviates from today's standard of 440 Hz (A4). In summary, these deviations in timbre, tuning, and recording noise lead to a so-called "domain shift", i.e., the FG58 and MB97 recordings are not close enough to the transcriber's training data. As a result, the model's predictions are highly unstable and do often not correspond to the actually played notes.

Fine-tuning the transcriber. Despite the aforementioned issues, our goal is to obtain highly accurate alignments on the BPSD. Therefore, we choose to adapt the T1 transcriber to the BPSD's audio versions by fine-tuning the model on the target data itself. Note that this is a valid procedure for the purpose of this study, as we do not evaluate the transcription accuracy itself, and we only use unaligned pairs of audio and score data for finetuning (see [4] for a detailed description of the training process using unaligned pairs of audio and score data). Therefore, we do not overfit the model towards a reference alignment. We denote the fine-tuned transcriber as T3.

Version	FB_RAW_OF		T1_RAW_OF		T3_RAW_OF	
	med.	cf 95.	med.	cf. 95	med.	cf. 95
AB96	19	132	11	58	12	58
AS35	20	157	11	62	12	76
DB84	20	149	12	71	14	93
FG58	19	137	27	*788*	12	80
FG67	20	138	10	49	11	59
FJ62	22	185	11	80	13	84
JJ90	17	102	10	56	12	56
MB97	23	217	12	*146*	12	103
MC22	22	178	13	80	14	89
VA81	19	143	11	61	12	69
WK64	16	99	10	49	11	45
average	20	128	**12**	121	**12**	**62**

Table 5: Median and 95 % confidence interval of the absolute synchronization error for individual performances in the BPSD. All experiments use raw frame and onset features (RAW_OF). Values are given in milliseconds.

Results with fine-tuned transcriber. After fine-tuning on the BPSD, the T3 model significantly improves the 95 % confidence interval for the two problematic versions FG58 and MB97 from 788 ms to 80 ms and from 146 ms to 103 ms, respectively. For all other audio versions, the median absolute error and the 95 % confidence interval of the fine-tuned transcriber T3 remain in a similar range as the original model T1. We note that the averaged 95 % confidence interval of 62 ms for the fine-tuned transcriber T3 is in the range of the typical tolerance in beat-tracking applications (70 ms), making the proposed synchronization approach with raw features even useful for the creation of datasets with high demands regarding timing.

6. CONCLUSION AND OUTLOOK

In this paper, we analyzed audio synchronization using raw features from transcription models. By conducting quantitative analysis on two different datasets of piano music, we show that the amount of alignment outliers is vastly reduced when using raw instead of symbolic features. We put a particular emphasis on the analysis of synchronization robustness of real-world audio recordings including historic instruments and recordings of low quality, and outline which acoustic conditions lead to alignment mismatch. By fine-tuning a transcription model on the target dataset and using the predicted raw features, we achieve synchronization accuracy that enables usage of the datasets even in time-critical applications such as beat tracking. As the raw features are computed anyway when using transcription models, we propose to use these raw features by default in synchronization pipelines. While raw features from transcription models yield excellent synchronization robustness for piano music, a yet unanswered question that we plan to address in future work is the performance in other genres, e.g., vocal or orchestral music.

7. ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 521420645 (MU 2686/17-1) and Grant No. 500643750 (MU 2686/15-1). 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] D. Schwarz, N. Orio, and N. Schnell, "Robust polyphonic midi score following with hidden Markov models," in *International Computer Music Conference (ICMC)*, Miami, Florida, USA, 2004.

[2] A. Arzt, G. Widmer, and S. Dixon, "Automatic page turning for musicians via real-time machine listening," in *ECAI 2008 - 18th European Conference on Artificial Intelligence, Patras, Greece, July 21-25, 2008, Proceedings*, ser. Frontiers in Artificial Intelligence and Applications, vol. 178. IOS Press, 2008, pp. 241–245.

[3] S. Ewert, B. Pardo, M. Müller, and M. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, April 2014.

[4] B. Maman and A. H. Bermano, "Unaligned supervision for automatic music transcription in the wild," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2022, pp. 14 918–14 934.

[5] F. Zalkow, S. Rosenzweig, J. Graulich, L. Dietz, E. M. Lemnaouar, and M. Müller, "A web-based interface for score following and track switching in choral music," in *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018.

[6] C. Dittmar, B. Lehner, T. Prätzlich, M. Müller, and G. Widmer, "Cross-version singing voice detection in classical opera recordings," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, October 2015, pp. 618–624.

[7] R. B. Dannenberg, "An on-line algorithm for real-time accompaniment," in *Proceedings of the International Computer Music Conference (ICMC)*, Paris, France, 1984, pp. 193–198.

[8] Y. Özer, S. Schwär, V. Arifi-Müller, J. Lawrence, E. Sen, and M. Müller, "Piano concerto dataset (PCD): A multitrack dataset of piano concertos," *Transaction of the International Society for Music Information Retrieval (TISMIR)*, vol. 6, no. 1, pp. 75–88, 2023.

[9] C. Weiß, F. Zalkow, V. Arifi-Müller, M. Müller, H. V. Koops, A. Volk, and H. Grohganz, "Schubert Winterreise dataset: A multimodal scenario for music analysis," *ACM Journal on Computing and Cultural Heritage (JOCCH)*, vol. 14, no. 2, pp. 25:1–18, 2021.

[10] C. Weiß, V. Arifi-Müller, M. Krause, F. Zalkow, S. Klauk, R. Kleinertz, and M. Müller, "Wagner Ring Dataset: A complex opera scenario for music processing and computational musicology," *Transaction of the International Society for Music Information Retrieval (TISMIR)*, vol. 6, no. 1, pp. 135–149, 2023.

[11] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 1869–1872.

[12] M. Müller, *Fundamentals of Music Processing – Audio, Analysis, Algorithms, Applications*. Springer Verlag, 2015.

[13] C. Schörkhuber and A. P. Klapuri, "Constant-Q transform toolbox for music processing," in *Proceedings of the Sound and Music Computing Conference (SMC)*, Barcelona, Spain, 2010.

[14] M. Müller, Y. Özer, M. Krause, T. Prätzlich, and J. Driedger, "Sync Toolbox: A Python package for efficient, robust, and accurate music synchronization," *Journal of Open Source Software (JOSS)*, vol. 6, no. 64, pp. 3434:1–4, 2021.

[15] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, "On the potential of simple framewise approaches to piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, New York City, New York, USA, 2016, pp. 475–481.

[16] J. Thickstun, Z. Harchaoui, and S. M. Kakade, "Learning features of music from scratch," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[17] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, New York City, New York, USA, 2016, pp. 37–43.

[18] J. Thickstun, Z. Harchaoui, D. P. Foster, and S. M. Kakade, "Invariances and data augmentation for supervised music transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 2241–2245.

[19] Y. Wu, B. Chen, and L. Su, "Polyphonic music transcription with semantic segmentation," in *Proceedings of the IEEE International Conference on Acoustics,*

Speech, and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 166–170.

[20] C. Weiß, J. Zeitler, T. Zunner, F. Schuberth, and M. Müller, "Learning pitch-class representations from score–audio pairs of classical music," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Online, 2021, pp. 746–753.

[21] J. Schlüter and S. Böck, "Improved musical onset detection with convolutional neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 6979–6983.

[22] S. Böck, A. Arzt, F. Krebs, and M. Schedl, "Online real-time onset detection with recurrent neural networks," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, York, UK, September 2012.

[23] S. Böck, F. Krebs, and G. Widmer, "Joint beat and downbeat tracking with recurrent neural networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, New York City, New York, USA, 2016, pp. 255–261.

[24] Y. Özer, M. Istvanek, V. Arifi-Müller, and M. Müller, "Using activation functions for improving measure-level audio synchronization," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, 2022, pp. 749–756.

[25] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C. A. Huang, S. Dieleman, E. Elsen, J. H. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, Louisiana, USA, 2019.

[26] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. H. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference, (ISMIR)*, Paris, France, 2018, pp. 50–57.

[27] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3707–3717, 2021.

[28] J. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. H. Engel, "MT3: multi-task multitrack music transcription," *Computing Research Repository (CoRR)*, vol. abs/2111.03017, 2021.

[29] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.

[30] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. H. Engel, "Sequence-to-sequence piano transcription with transformers," pp. 246–253, 2021.

[31] F. Foscarin, A. McLeod, P. Rigaux, F. Jacquemard, and M. Sakai, "ASAP: a dataset of aligned scores and performances for piano transcription," in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, 2020, pp. 534–541.

[32] X. Riley, D. Edwards, and S. Dixon, "High resolution guitar transcription via domain adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seoul, South Korea, 2024, pp. 1051–1055.

[33] T. Prätzlich, J. Driedger, and M. Müller, "Memory-restricted multiscale dynamic time warping," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, March 2016, pp. 569–573.

[34] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[35] J. Zeitler, C. Weiß, V. Arifi-Müller, and M. Müller, "BPSD: A coherent multi-version dataset for analyzing the first movements of Beethoven's piano sonatas." *Transactions of the International Society for Music Information Retrieval (TISMIR)*, submitted 2024.

[36] D. Edwards, S. Dixon, E. Benetos, A. Maezawa, and Y. Kusaka, "A data-driven analysis of robust automatic piano transcription," *IEEE Signal Process. Lett.*, vol. 31, pp. 681–685, 2024.

[37] T. Prätzlich and M. Müller, "Triple-based analysis of music alignments without the need of ground-truth annotations," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, March 2016, pp. 266–270.