Source	Datasets	Instruments
[16], [17], [2]	The Vienna 4x22 Piano Corpus [18]	Piano
[11], [8], [1]	Bach10 Dataset [19]	Violin, Clarinet, Tenor Sax, Bassoon
[8], [20], [21]	RWC Database [22]	Polyphonic Multi Instrument
[8]	28 mins of Amateur practice	Clarinet
[7], [2], [4]	MUS Subset of MAPS Database [23]	Piano
[2]	Mozart Sonatas [24], Rachnmaninoff Prelude Op. 23 No. 5 [25]	Piano
[3]	Synthetic dataset based on MSMD [26] and private data	Piano

Table 1: Datasets used across audio to score alignment research

2.3 Evaluation

Cont et al. [35] formalize the quantitative performance metrics of Score Following, forming the basis of the MIREX challenge for the task. Only a subset of their metrics are relevant for ASA since there is no expectation of real-time execution. They define the error as $\mathrm{e}_i = \mathrm{t}_i^e - \mathrm{t}_i^r$, where t_i^e is the estimated time in the score for event i, and t_i^r is the actual time of event i in the reference. The alignment corresponding to an event is considered misaligned only if it exceeds a time threshold θ_e, which we call the misalignment threshold, noting that events could be notes or other time references (See Thickstun et al. [5] for a discussion on the difference between temporal and note based metrics). Accordingly, the following metrics are proposed: the Standard Deviation of the e_i of non misaligned events, the Misalignment Rate (MR) (percentage of events with $|\mathrm{e}_i| \geq \theta_e$, and the Average Imprecision (average absolute error of non misaligned events). For system-wide metrics, they propose the Piece-wise Precision Rate (PPR) over a related subset of scores, calculated as the percentage of non misaligned notes, and the overall precision rate (OPR) calculated similarly to the PPR but over the whole database instead. In practice, researchers slightly vary in their evaluation metrics. They mostly capture the Alignment Rate (AR), according to a set of θ_e usually between 50 ms to 300 ms, sometimes using it as an analogue for PPR.

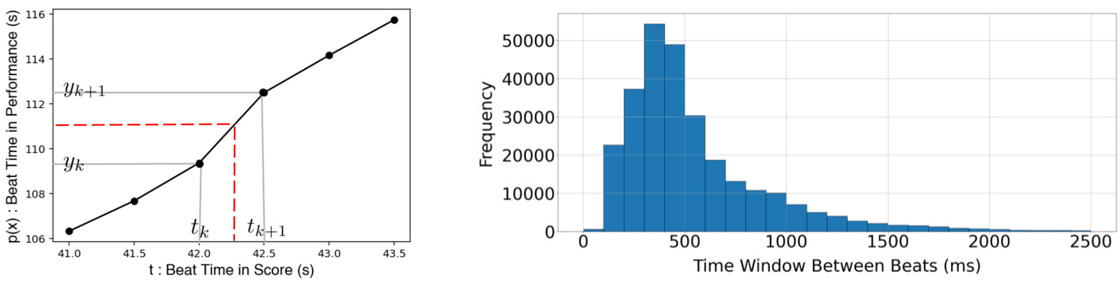
Another commonly used metric is the Average Alignment Error (AAE) defined in [11], which is the average absolute error for each audio frame, distinguishing it from Average Imprecision, which is calculated for non misaligned events only. AAE can be reported in milliseconds or in beats, depending on the end goal in mind [11]. Without using AAE explicitly, Jiang et al. [12] calculate the proportion of frames misaligned by unites expressed in beats per measure. A metric with the same essence as AAE is used in [2], where they additionally report the median, 1st, and 3rd quartiles of this difference. In addition to the aforementioned metrics, some authors conduct an extent of qualitative analysis in order to make useful insights about their systems with respect to the scope in which the problem is defined. This has been done to test robustness to performance mistakes [8], for error prone scores [20], to understand the impact of polyphony [11, 19, 20], or the presence of percussion [20], tempo variations [2, 19], or skips [8, 10, 12].

3. GENERATING GROUND TRUTHS FROM ASAP

Reusing the ASAP dataset [15] is a reasonable step to expand the data available for ASA research. First, it offers aligned audio and music scores on a beat level for 520 piano performances over various composers and styles. Some pieces are performed by several pianists, and we observe temporal variation in the different interpretations of one piece. In addition, we believe that these alignments could help us create more data by introducing structural variations within a single piece or across different pieces, depending on the scope of the alignment problem we want to consider, which we highlight in Section 6 along with other augmentation ideas to cover a variety of ASA problems. The rest of this section describes how we create ASA ground truth approximations from the beat annotations of ASAP, along with the potential implications and pitfalls of doing so.

3.1 From beat annotations to full alignments

We use the aligned beat annotations of the performance MIDIs and score MIDIs provided by the ASAP dataset to obtain approximated ground truth alignments (performance-aligned scores) for score-performance pairs at a low cost through Piecewise Linear interpolation. Every beat in the score is mapped to a specific time in the performance, giving an alignment function with which we map each onset time of the MIDI score file to a time in the performance audio. A schematic is shown in Fig 1a. This approach does not give an alignment with a note-to-note resolution. However, we believe it is still usable for evaluating methods outputting temporal alignments that inherently do not provide this level of precision, such as warping paths obtained by DTW alignments, or for training audio to score alignment systems with methods that tolerate weakness in the reference alignments. To understand the extent of the error, we investigate the temporal resolution of the beat annotations (the time distances between consecutive beats) over the chosen subset of the ASAP dataset. As shown in Fig. 1b), the majority of such distances fall between 200 and 1100 ms. Clearly, the faster the tempo of the performance, the less spaced in time consecutive beat annotations are. For context, the distance between two quarter notes in a 120 BPM score is 500 ms.



(a) (b)

Figure 1: (a) Illustrative snippet of the ground truth alignment from the GuoE01M Prelude BMV883 performance. (b) Distributions of the time in between consecutive beats for all files in our chosen subset of the ASAP dataset.

3.2 Error Limits

To understand the limits of the error between the ideal ground truth and the approximated ground truth (ϵ_{gt}) we resort to the definition of Piecewise Linear Interpolation, shown as follows:

$$p(x) = y_k + \frac{y_{k+1} - y_k}{t_{k+1} - t_k}(x - t_k), \text{ for } x \in [t_k, t_{k+1}], \quad (1)$$

where x is the time point in the score for which we need to approximate a corresponding time in the performance, t are reference points in the score annotation (in the context of the ASAP dataset, these would be beat times), and y are time reference points in the corresponding performance annotation. This is demonstrated in Fig 1a. In reality, the path between (t_k, y_k) and (t_{k+1}, y_{k+1}) can take any shape as long as it is monotonic (an assumption we can make due to our data). Taking the extreme unrealistic case where $p(x)$ takes on either y_{k+1} or y_k, the ground truth approximation error ϵ_{gt} must be:

$$\epsilon_{gt} = \max(|p(x) - y_k|, |p(x) - y_{k+1}|), x \in [t_k, t_{k+1}]. \quad (2)$$

If x falls on the midpoint of t_k and t_{k+1}, then ϵ_{gt} cannot surpass $\frac{1}{2}(y_k, y_{k+1})$, meaning that even for beats highly spaced apart (1000 ms) the error would be 500ms. Moreover, we would argue that in practice the error would be even less, because of the musical flow. However, the potential of error is a limitation due to which a decision needs to be made on which files to discard. Although not much detail is provided, it is worth noting that Duan and Pardo [19] mention their use of beat annotations to create ground truths, meaning that this process has been accepted in past studies, although it was not discussed elaborately.

4. DATA VALIDATION

The approximated ground truths of our interpolated dataset can have two problem sources: 1) misalignments within the generated annotations due to the low resolution of the score or performance beat annotations (as discussed in Section 3.2), and 2) annotation problems from the original dataset. These need investigation before using the dataset, whether for evaluation or training. We create test audio for every generated annotation file, where the left channel includes the performance-aligned score (the approximated ground truth) and the performance audio on the right channel. Therefore, by listening to all such test audios, it is possible to get a sense of both problems above. However, due to the size of the ASAP subset we reuse, it was not feasible to listen to all the hours of audio. So, we conduct a typical ASA experiment to help give clues as to which files most likely could contain errors and therefore would need to be examined. The rationale of this selective investigation is that files with low misalignment rates have passed an implicit check. Suppose a music score has been aligned with the performance audio using a process different from that with which the ASAP dataset was created. If this result matched the interpolated ground truth, then it is improbable that there is a problem with its beat alignment. Otherwise, there would have been a difference between the alignment result and the ground truth. Moreover, to determine whether we should discard files with large time windows between beats, we listened to a selection of the files with intra-beat annotation time differences of $\approx$ 1500 ms. They sounded correct for several performances, especially for files performed without much temporal variation (such as the Fugue BMV 874 and Prelude BMV 863 performances by Kurz and Shyc, respectively). We decided to keep all such files and only discard them based on the results of the alignment experiment.

4.1 Alignment Experiment

To align a performance and its symbolic score, we sonify the latter using the fluidsynth[2] and conduct DTW-based audio to audio alignment between the synthesized score and the audio performance. We use the librosa[3] [36] DTW implementation, and the distance matrix is computed by applying the Euclidean distance between the feature vectors.

[2] https://www.fluidsynth.org/
[3] https://librosa.org/doc/main/generated/librosa.sequence.dtw.html

4.1.1 *Features*

We use the CQT and 5 of the chroma representations compared in [37], which are: a Connectionist Temporal Classification loss trained chroma extractor (CTC-Chroma) explained in [38]; Non Negative Least Squares (NNLS) chroma [39]; the Harmonic Pitch Class Profile (HPCP) [40]; the Deep Chroma Extractor (DCE) [41]; and the classic Chroma algorithm implemented in [36]. All these algorithms are easily usable out of the box, and we believe are appropriate to use for piano data. Details on the parameters of each algorithm can be found in the accompanying repository [4].

4.1.2 *Quantitative Results*

Since our focus is on filtering files, we report file-based metrics rather than global metrics for each DTW system. Therefore we omit reporting the Overall Alignment Rate (OAR), the system-wide AR considering the notes over all scores (or all temporal units). We use the Average Absolute Error (AAE) for each file in the dataset, shown in the box and whisker plots of Fig. 2 indicating the 1st, median, and 3rd quartiles per each DTW system. We also maintain the AR and the Absolute Errors (AE) for each file. We use these metrics to identify 1) suspects of files with beat alignments that are not highly accurate, which should be discarded, or 2) files for which our ground truth approximation approach yielded a high ϵ_{gt} for annotations within the beats. Those might still be relevant to keep depending on how they will be used. We explain this process in Section 5. However, we must be careful before discarding any files based on performance metrics to avoid cherry-picking only the files for which our ASA systems perform well. This is why we use a variety of audio representations, knowing that some might not be perfectly suitable for audio to score alignment, and no files are discarded unless they performed badly using all DTW systems.

5. RESULT INFORMED DATA FILTERING

Although not the core of our work, we observe that the DTW systems using CQT, HPCP, and Chroma perform better than the rest. This can be seen from Fig. 2 from the lower AAE time windows and the compactness of their distribution, and although we do not show a plot for OAR, these three systems reach very high OARs within the 0 - 60ms error thresholds. In fairness, the CQT system with the best performance is the gold standard obtained by the Bayesian Optimization of [14], suggesting that before making any absolute statements about the superiority of any of the DTW systems, their parameters should be optimized similarly. Besides, comparing such systems or finding the best performing ASA system is not the goal of this paper, and as we argue earlier, ASA is still missing a clear methodology and varied data with which qualitative evaluation can be conducted. This hinders the ability to compare between systems. The ASA results shown are just a means to an end, which is validating the interpolated dataset as described in Section 4, and helping us pinpoint problems and the potential need to filter some files, as shown in Sections 5.1 and 5.2.

[4] `https://github.com/Alia-morsi/asa_benchmarks`

5.1 AAE based investigation

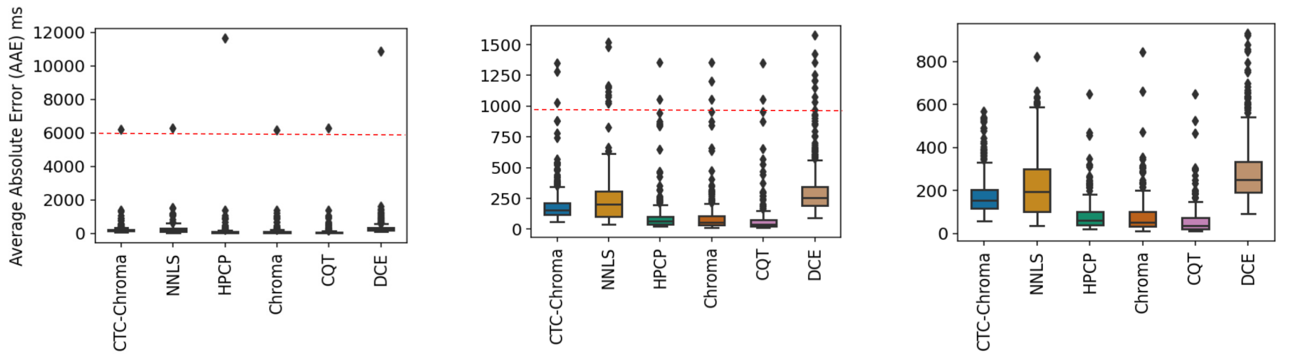
Without discarding any generated alignments yet, we start by observing the box-and-whisker plots showing the AAE for all the 520 usable files of ASAP evaluated over the interpolated ground truth references, shown in Fig. 2a. We observe files with very high AAEs, most likely signalling either an annotation or calculation error since they represent alignment error values that are unreasonably high. Drawing a threshold at an AAE of 6000 ms (the red line) allows us to filter those clear outliers, thus arriving at the second plot shown in Fig. 2b. Then, we decide to conduct further filtering at a threshold at an AAE of 1000, arriving at Fig. 2c. We can keep setting lower AAE threshold and filtering more files for as long as needed. But the idea is to listen to the test audio described in Section 4 and to observe the annotations before discarding a file, to make sure that we are not filtering good ground truth approximations.

5.2 AR based investigation

Files for which the AR is very low (approx 10%) at θ_e thresholds between 50 and 100 ms signal the need for further investigation. In Section 4 we referred to two problems: 1) the possibility of a temporal offset in the ground truth annotation of the original ASAP dataset, and 2) the possibility of the generated labels being misaligned due to large temporal distances between consecutive beat annotations. If for a file we observe that the 1st quartile, median, and 3rd quartiles do not progress ascendingly as expected a (eg. if the 3 values are nearly equal) and are higher than usual, then this could indicate a temporal offset in the ground truth annotation. Through the listening verification we describe in Section 3, we found that this is the case for at least 6 score-performance pairs. As for the latter problem, if we find that if a file has a low AR, and the 1st, median, and 3rd quartiles of the AE move ascendingly (as expected), then it is a suspect of the low resolution problem. Examples of such files are the 2 performances of Prelude BMV 846, and Prelude BMV 867, where it is clear upon listening that there is a high temporal variation at a phrase level. When coupled with an insufficient resolution of the beat annotations, this would certainly cause ground truth errors. Files of the first category should always be discarded, but files of the second category could be kept, depending on the extent of the misalignment introduced through the ground truth approximation, and the tolerance allowable by the expected use.

5.3 Limitations

A legitimate criticism of our work would be that ground truths generated from the ASAP dataset do not live up to the ambition driving the paper, which is to create a large benchmark for ASA research covering a variety of musical use-cases. Although we do not fully dispute this because



(a) Alignment results for all MIDI score-audio pairs. Next cutoff threshold = AAE 6000 ms.

(b) Alignment results for MIDI score-audio pairs with AAE < 6000 ms. Next cutoff threshold = 1000 ms.

(c) Alignment results for MIDI score-audio pairs with AAE < 1000 ms. We do not apply further filtering

Figure 2: Box and Whisker plots showing the Average Alignment Error (AAE) results of the DTW systems using each of the chroma extraction algorithms, calculated using the approximated ground truths. Lower results are better. The red horizontal line of a figure indicates the cutoff threshold to be applied for generating the figure to its right.

all the scores of the ASAP dataset are monotonically increasing classical solo piano performances which highly adhere to their music scores in the performance, our point is that neither the ASAP dataset nor any other single accessible dataset would possess the level of diversity needed to move past the bottleneck. The goal is to start accumulating adapted datasets to eventually arrive at a bigger benchmark. For example, a similar process could be applied to the MazurkaBL dataset [42], and perhaps several others too, although the data preparation methodology and corresponding discussion are coupled with the specifics of the chosen dataset. Moreover, as we better explain in Section 6, even with the generated ground truth approximations from ASAP alone, there is room to create interesting data extensions with the approximated beat alignments. Another drawback of our work could be that the results informed data investigation described in Section 5 is not enough, and there should be a more rigorous manual verification process of the derived ground truths. We agree that manual verification of the whole datatset would be ideal, but we also defend that finding compromises for practical benefit should not be disregarded while being very clear on where these datasets fail. Moreover, perhaps a confidence measure can be created based on comparing the correlation of onsets between the left and right channels of the test audio described in Section 4. Further limitations of this work are that we do not discuss the computational complexity of most ASA methods, and rather constrain the use of the term bottleneck to conceptual hindrances facing ASA.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we argue that ASA research has reached glass ceiling, and a crucial way to get past it is to unify what would be considered core to the problem definition in terms of musical scenarios. For example, what kinds of structural variations between the audio and score should be considered, what kinds of instruments should be supported, what recording quality is expected, etc. We believe this would not be possible without developing benchmarks covering such scenarios, which would support a paradigm shift in how ASA is approached, and would allow us to compare between the performance of ASA systems developed by different researchers. To take a first step towards increasing the size and variety of data, we demonstrate the reuse of the 520 scores of ASAP dataset for which beat aligned scores and performance audio pairs are available. We argue that despite its creation with other MIR research topics in mind, it still can be a very useful resource for researchers interested in ASA for classical piano music. We conduct several data validation steps informed by the AAE and AR from results from a classic DTW pipeline, allowing a selective investigation and filtering of the dataset.

6.1 Future Directions

In addition to adapting more related datasets, we would like to build on this work by artificially extending the data to improve its balance. We need to include cases where the audio performance does not adhere well to the music score, whether through skips, repeats, or performance mistakes. Starting from the generated alignment ground truths (or alignment references from other datasets) we could create semi-artificial data where we shuffle parts of the score, and concatenate the audio from the real performance to match this modified score. To avoid these modifications sounding unnatural, we could try and choose realistic parts of the piece, referring to works on music structure analysis to introduce structural repetitions with more musical sense. Datasets with structural variations would be interesting especially to improve the ability of ASA systems to recover when lost, which is relevant for real-time audio to score alignment. Nevertheless, covering a wider range of instruments still poses a challenge, but this is expected to become easier as synthesis technologies develop further. Finally, we conclude with our hope that ASA approaches would find more inspiration from recent advances in Cover Song Detection and Natural Language Processing.

7. ACKNOWLEDGEMENTS

This research was carried out under the project Musical AI - PID2019-111403GB-I00/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación.

8. REFERENCES

[1] J. Carabias, F. Rodríguez-Serrano, P. Vera-Candeas, N. Ruiz Reyes, and F. Canadas Quesada, "An audio to score alignment framework using spectral factorization and dynamic time warping," in *Proceedings of the 16th International Society for Music information Retrieval Conference (ISMIR)*, October 2015, pp. 742–748.

[2] A. Arzt and S. Lattner, "Audio-to-score alignment using transposition-invariant features," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, September 2018, pp. 592–599.

[3] R. Agrawal, D. Wolff, and S. Dixon, "Structure-aware audio-to-score alignment using progressively dilated convolutional neural networks," in *Proceedings of the 47th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 571–575.

[4] R. Agrawal and S. Dixon, "Learning frame similarity using siamese networks for audio-to-score alignment," in *Proceedings of the 28th European Signal Processing Conference (EUSIPCO)*, January 2020, pp. 141–145.

[5] J. Thickstun, J. Brennan, and H. Verma, "Rethinking evaluation methodology for audio-to-score alignment," *arXiv preprint arXiv:2009.14374*, 2020.

[6] F. Simonetta, S. Ntalampiras, and F. Avanzini, "Audio-to-score alignment using deep automatic music transcription," in *Proceddings of the 23rd IEEE International Workshop on Multimedia Signal Processing (MMSP)*, October 2021.

[7] T. Kwon, D. Jeong, and J. Nam, "Audio-to-score alignment of piano music using rnn-based automatic music transcription," in *Proceedings of the 14th Sound and Music Computing Conference (SMC)*, July 2017, pp. 380–385.

[8] T. Nakamura, E. Nakamura, and S. Sagayama, "Real-time audio-to-score alignment of music performances containing errors and arbitrary repeats and skips," *IEEE/ACM Transactions of Audio, Speech and Language Processing*, vol. 24, no. 2, pp. 329–339, February 2016.

[9] M. Dorfer, A. Arzt, and G. Widmer, "Learning audio-sheet music correspondences for score identification and offline alignment," in *Proceedings of the 18th International Society for Music Information Retreival Conference (ISMIR)*, October 2017, pp. 115–122.

[10] M. Shan and T. J. Tsai, "Automatic generation of piano score following videos," *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 4, p. 29–41, March 2021.

[11] Z. Duan and B. Pardo, "A state space model for online polyphonic audio-score alignment," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 197–200.

[12] Y. Jiang, F. Ryan, D. Cartledge, and C. Raphael, "Offline score alignment for realistic music practice," in *Proceedings of the 16th Sound and Music Computing Conference (SMC)*, May 2019, pp. 387–393.

[13] E. Nakamura, K. Yoshi, and H. Katayose, "Performance error detection and post-processing for fast and accurate symbolic music alignment," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, October 2017, pp. 347–353.

[14] C. Raffel and D. P. W. Ellis, "Optimizing DTW-based audio-to-midi alignment and matching," in *41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 81–85.

[15] 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 21st International Society for Music Information Retrieval Conference (ISMIR)*, October 2020, pp. 534–541.

[16] S. Dixon, "An on-line time warping algorithm for tracking musical performances," in *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, August 2005, pp. 1727–1728.

[17] A. Arzt, G. Widmer, and S. Dixon, "Automatic page turning for musicians via real-time machine listening," in *Proceedings of the 18th European Conference on AI (ECAI)*, July 2008, pp. 241–245.

[18] W. Goebl. (1999) The vienna 4x22 piano corpus. http://dx.doi.org/10.21939/4X22.

[19] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, pp. 1205–1215, October 2011.

[20] C. Joder, S. Essid, and G. Richard, "A comparative study of tonal acoustic features for a symbolic level music-to-score alignment," in *Proceedings of the 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 409–412.

[21] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features,"

in *Proceedings of the 34th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2009, pp. 1869–1872.

[22] M. Goto, "RWC music database: Popular, classical, and jazz music databases," *Proceedings of the 3rd International Society for Music Information Retrieval Conference (ISMIR)*, vol. 1, pp. 287–288, October 2002.

[23] 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, August 2010.

[24] G. Widmer, "Discovering simple rules in complex data: A meta-learning algorithm and some surprising musical discoveries," *Journal of Artificial Intelligence*, vol. 146, no. 2, pp. 129–148, June 2003.

[25] A. Arzt, "Flexible and robust music tracking," Ph.D. dissertation, Department of Computational Perception, Johaness Kepler University, Linz, December 2016.

[26] M. Dorfer, A. Hajič jr. J.and Arzt, H. Frostel, and G. Widmer, "Learning audio–sheet music correspondences for cross-modal retrieval and piece identification," *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 1, p. 22–33, September 2018.

[27] F. Henkel, S. Balke, M. Dorfer, and G. Widmer, "Score following as a multi-modal reinforcement learning problem," *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 2, pp. 67–81, November 2019.

[28] S. Salvador and P. K.-F. Chan, "Toward accurate dynamic time warping in linear time and space," in *Journal of Intelligent Data Analysis*, vol. 11, no. 5, October 2007, pp. 561–580.

[29] T. Prätzlich, J. Driedger, and M. Müller, "Memory-restricted multiscale dynamic time warping," in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 569–573.

[30] 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*, vol. 6, no. 64, p. 3434, August 2021.

[31] C. Fremerey, M. Müller, and M. Clausen, "Handling repeats and jumps in score-performance synchronization," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, August 2010, pp. 243–248.

[32] M. Grachten, M. Gasser, A. Arzt, and G. Widmer, "Automatic alignment of music performances with structural differences," in *Proceedings of the 14th International Society for Music Information Retreival Conference (ISMIR)*, November 2013, pp. 607–612.

[33] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*, 1st ed. Springer Publishing Company, Incorporated, 2015.

[34] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. Bello, "Deep salience representations for F0 estimation in polyphonic music," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, October 2017, pp. 63–70.

[35] A. Cont, D. Schwarz, N. Schnell, and C. Raphael, "Evaluation of real-time audio-to-score alignment," in *Proceedings of the 8th International Society for Music Information Retreival Conference (ISMIR)*, September 2007, pp. 315–316.

[36] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proc. 14th Python in Science Conference (SciPy)*, July 2015, pp. 18–24.

[37] M. P. Fernández, H. Kirchhoff, and X. Serra", "A comparison of pitch chroma extraction algorithms," in *Proceedings of the Sound and Music Computing Conference*, Saint Etienne, France, June 2022, pp. 222–229.

[38] C. Weiss and G. Peeters, "Training deep pitch-class representations with a multi-label CTC loss," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, Virtual Event, November 2021, pp. 754–761.

[39] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, August 2010, pp. 135–140.

[40] E. Gómez, "Tonal description of polyphonic audio for music content processing," *INFORMS Journal on Computing*, vol. 18, no. 3, pp. 294–304, 08 2006.

[41] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," in *Proceedings of the 17th Intternational Society for Music Information Retreival Conference (ISMIR)*, August 2016, pp. 37–43.

[42] K. Kosta, O. F. Bandtlow, and E. Chew, "MazurkaBL: Score-aligned loudness, beat, expressive markings data for 2000 chopin mazurka recordings," in *Proceedings of the 4th International Conference on Technologies for Music Notation and Representation (TENOR)*, May 2018, pp. 85–94.

Papers