ASAP: a dataset of aligned scores and performances for piano transcription

Francesco Foscarin, Andrew Mcleod, Philippe Rigaux, Florent Jacquemard, Masahiko Sakai

ASAP: A DATASET OF ALIGNED SCORES AND PERFORMANCES FOR PIANO TRANSCRIPTION

Francesco Foscarin[1] **Andrew McLeod**[2] **Philippe Rigaux**[1]
Florent Jacquemard[1,3] **Masahiko Sakai**[3]
[1] CNAM, Paris, France [2] EPFL, Lausanne, Switzerland
[3] INRIA, Paris, France [4] Nagoya University, Nagoya, Japan
francesco.foscarin@cnam.fr, andrew.mcleod@epfl.ch, philippe.rigaux@cnam.fr
florent.jacquemard@inria.fr, sakai@i.nagoya-u.ac.jp

ABSTRACT

In this paper we present Aligned Scores and Performances (ASAP): a new dataset of 222 digital musical scores aligned with 1068 performances (more than 92 hours) of Western classical piano music. The scores are provided as paired MusicXML files and quantized MIDI files, and the performances as paired MIDI files and partially as audio recordings. Scores and performances are aligned with downbeat, beat, time signature, and key signature annotations. ASAP has been obtained thanks to a new annotation workflow that combines score analysis and alignment algorithms, with the goal of reducing the time for manual annotation. The dataset itself is, to our knowledge, the largest that includes an alignment of music scores to MIDI and audio performance data. As such, it is a useful resource for a wide variety of MIR applications, from those that target the complete audio-to-score Automatic Music Transcription task, to others that target more specific aspects (e.g., key signature estimation and beat or downbeat tracking from both MIDI and audio representations).

1. INTRODUCTION

As data-hungry deep learning models have become more ubiquitous in the field of MIR in recent years (e.g., [19, 22, 37]), large, well-annotated datasets have become increasingly important. Similar trends towards deep learning methods have been seen in related fields such as natural language processing [42] and computer vision [39]. For many tasks in these fields, large datasets can be automatically scraped from the web, and annotated quickly by non-experts (e.g., [34]). Unfortunately, the same can often not be said for tasks in MIR, for many reasons.

First, producing high-quality audio or MIDI data is a non-trivial task, requiring expert performers and expensive equipment that is not always available. Second, the availability of a high-quality digital ground truth is not guaranteed in many cases, particularly for tasks which require a musical score [1], e.g., for the complete audio-to-score Automatic Music Transcription (AMT) task (see [1] for a recent overview of AMT). Finally, even in the case when both audio/MIDI data and high-quality digital ground truth scores are available, acquiring an alignment between the two is non-trivial. Automatic alignment methods (e.g., [30]) are often not robust enough to deliver highly reliable results and require time-consuming post-processing and cleaning by expert musicians for advanced data usage.

We introduce the Aligned Scores and Performances (ASAP) dataset [2], containing digital musical scores of Western classical piano pieces as both MusicXML and MIDI, aligned at the beat level with audio (from MAESTRO [17]) and MIDI recordings of over 1000 human performances. Regarding the three difficulties outlined above (data creation, digital ground truth availability, and recording-ground truth alignment), we use (1) publicly available MIDI and audio from expert human performances; (2) publicly-available musical scores scraped from the web; and (3) a new workflow to efficiently produce aligned beat, downbeat, key, and time signature annotations with minimal human correction required.

Although ASAP can be used for many tasks, it was designed with two categories specifically in mind: (1) complete audio-to-score AMT and (2) metrical structure-based tasks (e.g., beat and downbeat tracking, tempo estimation, metrical structure alignment, and rhythm quantization) of classical piano performance, from both audio and MIDI.

While a major goal of AMT is to convert an input audio recording into a form of human-readable music notation, the vast majority of AMT systems fall short of such an output. Rather, they convert the input audio recording into some sort of time-frequency representation: either a frame-based multi-pitch detection, where the presence of each pitch is estimated at each point in the input recording (e.g., [21]); or a note-based output such as a piano-roll or MIDI file, where notes are detected each with a pitch, an onset time, and an offset time (e.g., [16]). [3] For these purposes, the ground truth only needs to contain some aligned

[1] Although PDF scores are sometimes available, and the field of Optical Music Recognition (see [3] for a recent overview) involves converting these into digital format, this conversion can add errors, and starting from a clean, digital score is generally better if available.

[2] `https://github.com/fosfrancesco/asap-dataset`

[3] This can be post-processed into a musical score, as in the pipeline approach of [29], but such pipelines tend to add noise at each step.

pitch or note presence data—not a full musical score. Existing datasets such as MAPS [8] and the larger MAESTRO [17] contain appropriate ground truths for this level of transcription (see Section 2).

In recent years, a few systems have been designed to output a more complete musical score directly (e.g., [5, 35]). While each of these works shows promise on this difficult task, neither includes time signatures or key signatures in their output. One [5] uses synthetically generated scores (and synthesized audio), and the other [35] uses audio synthesized from real scores. Since the eventual goal of AMT is a full transcription from human performance to musical score, a large dataset of non-synthetic human performances (containing the tempo deviations and timing intricacies of human performance, as well as real audio) is needed: synthetic data can be useful in the initial phase of model design, but human performance data is required for a truly reliable evaluation. ASAP provides such a dataset.

To our knowledge, no audio-to-score transcription system exists that requires fine-grained alignments between recordings and ground truths, perhaps due to a lack of available data. However, the use of data with even a coarse alignment has been shown to improve performance on the related task of monophonic note-based transcription [31], suggesting that the same should be true for audio-to-score AMT, if a large enough dataset of aligned recordings and scores was available. ASAP provides this time-alignment between the included recordings and ground truth scores.

Regarding metrical tasks, large audio datasets exist, especially for beat and downbeat tracking, enabling sophisticated systems to be trained (e.g., [2]). However, there is an absence of similarly-sized annotated datasets consisting of MIDI data, resulting in a noted lack of training data for metrical tasks from MIDI input (e.g., [25]). Even in the case of audio, much of the existing data is dance music or from other genres with relatively steady tempi compared to the classical piano music contained in ASAP (see Section 4.3 for an analysis of ASAP's tempo changes).

We produce beat and downbeat annotations for every performance in ASAP with a novel workflow that exploits the precise metrical structure information available in a musical score. Projecting this onto each corresponding performance guarantees a robust means to identify the beat and downbeat positions. Working with MIDI allows us to overcome many difficulties found in similar approaches using only audio data (e.g., [32]). The workflow allows us to drastically reduce the time required for manual annotation.

2. RELATED WORK

Some public datasets similar to ASAP—containing combinations of musical scores, MIDI performances, and audio recordings, for AMT and/or beat tracking—already exist. In this Section, we describe those existing datasets in comparison to ASAP, highlighting specifically where ASAP addresses their deficiencies. Table 1 contains a summary of the largest of these datasets in comparison to ASAP. We first describe those containing musical performances (useful for AMT), and follow that with a brief discussion of available datasets for metrical structure-based tasks.

2.1 Performance datasets

There are two datasets which contain multiple performances of many different pieces. The Vienna 4x22 Piano Corpus [11] consists of 22 different performances of each of 4 different pieces, in both audio and MIDI format, aligned to a metrical grid. The CHARM Chopin Mazurka Project[4] dataset contains many recordings of each of 49 different Mazurkas composed by Frédéric Chopin, although the original audio recordings are only referenced, and not available online (instead, many pre-calculated features are provided). While these datasets are valuable for investigating live performance deviations and comparisons between different performances of the same piece, they are not as useful for AMT, since they each consist of a small number of different pieces, leading to likely model overfitting (only 4 different pieces for Vienna 4x22, and only pieces by a single composer in the Mazurka dataset).

The Saarland Music Data (SMD) dataset [28] contains 50 synchronized audio and MIDI recordings of human performers playing a Disklavier piano. The files are not aligned with any musical scores or beat annotations, and the dataset's size is somewhat small compared to other similar datasets. Likely because of its size, SMD has has not been used for AMT in recent work to our knowledge.

CrestMuse PEDB [15] is a dataset based on audio recordings of multiple piano performances of around 100 unique pieces. However, the original audio recordings are not included. Rather, references to commercial CDs which can be purchased, and on which the recordings can be found are given. After a PDF application and pledge are filled out and submitted, access is granted to download the database in about a week. Provided in the dataset are MIDI files, whose notes have been hand-aligned to the referenced recordings; and digital musical scores in an XML-based format, to which the notes and beats of the MIDI files are aligned (using "deviation" XML tags in the score files). Since its initial release, some audio files have been added. However, these are different from the original score-aligned audio recordings, and in some cases are synthesized from MIDI performance. The difficulty of acquiring the audio recordings makes this database rarely used for audio-based tasks such as AMT.

The piano-midi dataset[5] contains 324 quantized MIDI files whose tempo curves have been manually altered with the goal of becoming more human-like. The MAPS dataset [8] contains 210 of these MIDI files (of 119 different pieces), without key and time signatures, each paired with an audio file—some synthesized and some actual recordings. A-MAPS [41] later augmented MAPS with MIDI files containing key signature and time signature annotations. Since the MIDI data is generated from tempo-varied quantized MIDI, rather than actual performance, the MIDI files and recordings do not contain all of the timing variance that would be present in real performance: note onsets and offsets which lie on the same beat in the original musical score also occur simultaneously in the corresponding

[4] http://www.mazurka.org.uk/

[5] www.piano-midi.de

	Size		Performance		Quantized		Annotations		
Dataset	Total	Unique	Audio	MIDI	MIDI	Score	Alignment	Metrical	Key
MAPS [8]	269	52	Pseudo†	Pseudo	✓		Full	✓ [41]	✓ [41]
CrestMuse-PEDB [15]	411	≈100	Partial†	✓	✓	✓	Full	✓	✓
SUPRA [36]	478	≈430	Pseudo†	✓					
MAESTRO [17]	1282	≈430	✓	✓					
GTZAN [38]	1000	1000	✓					✓ [24]	Global
Ballroom [12]	685	685	✓					Beat [23]	
Hainsworth [14]	222	222	✓					Beat	
SMC [18]	217	217	✓					Beat	
ASAP	1068	222	520	✓	✓	✓	Beat	Beat	✓

Table 1. An overview of the most relevant datasets for AMT (top section) and metrical tasks (middle section), compared to ASAP (bottom). Alignment refers to the level of alignment between the quantized and performance data. MAPS consists of pseudo-live performance (quantized MIDI with manually altered tempo curves). †MAPS, SUPRA (fully) and CrestMuse-PEDB (partially) include synthesized audio (not real recordings).

MIDI and audio files. In real performance, such events only rarely occur simultaneously. Rather, small timing deviations introduce gaps, overlaps, and other timing variation (see e.g. [26], Figure 4), which are therefore missing (along with ornamentation such as trills, as well as performance errors). Although these datasets contain perfectly-aligned ground truth annotations (which has made MAPS a standard for AMT evaluation since its release), their modest size and the fact that they are not real live performance are drawbacks that we hope to address with ASAP.

The SUPRA dataset [36] contains 478 MIDI files of around 430 different pieces generated from an archive of piano performances in the form of physical piano rolls. SUPRA also contains synthesized audio recordings of each MIDI file, and labels each with a composer and title, but provides no metrical alignment of the pieces.

The MAESTRO dataset [17] contains 1282 real performances of around 430 different pieces from the Yamaha piano e-competition [6]. Each performance is available as a MIDI file and an audio recording with a fine alignment of around 3 ms. Metadata are available for each performance, including the composer and title of each. MAESTRO's size, fine alignment with ground truth, and the fact that it is real performance have made it an excellent source of training and evaluation data for AMT from recording to piano-roll. However, MAESTRO does not contain any note-, beat-, or even piece-level alignment with digital musical scores, required for the complete audio-to-score AMT task (and which ASAP does contain).

2.2 Metrical structure datasets

For metrical structure-based tasks, from live performance MIDI data, annotated datasets from the previous section (particularly piano-midi and CrestMuse-PEDB) are typically used. However, they are relatively small (especially in terms of unique pieces), and piano-midi files in particular do not contain real live performance, as mentioned. For the same tasks from audio data, large annotated datasets exist, enabling sophisticated models to be designed and trained (e.g., [2]). The largest and most widely used (where audio files are publicly available, including at least beat annotations) are: GTZAN [38] (1000 audio recordings of various genres with beat, downbeat, and 8th-note annotations), Ballroom [12] (685 audio recordings of ballroom dancing music with beat and downbeat annotations), Hainsworth [14] (222 audio recordings of Western music), and SMC [18] (217 audio recordings of Western music, specifically selected to be difficult for beat tracking). However, the music contained in these datasets tend to have a much steadier tempo than those contained in ASAP (even SMC; see Section 4.3 for a comparison).

3. PRODUCING MUSIC ANNOTATIONS

Annotating a dataset the size of ASAP with ground truth for AMT and related problems such as beat tracking is a time consuming task that can, in principle, only be performed by expert musicians. This severely limits the ease with which one can produce a reliable and finely crafted dataset at the required scale.

For piano music, MIDI and audio performances can be automatically aligned if they are recorded at the same time using an acoustic piano fitted with the proper sensors such as a Disklavier. The main problem then becomes annotating the performances with metrical markings (such as beats and downbeats) and aligning those with a musical score. Holzapfel et al. [18] describe the process used to annotate the SMC dataset in detail, showing how much manual work was required for its beat annotations. ASAP contains more than 92 hours of performance data, and even a skilled musician would need to listen to each multiple times with the proper annotation software in order to annotate it fully (and this time can increase dramatically for complicated pieces with multiple time and key signature changes). Moreover, as highlighted in [40], the manual annotations would still be affected by human subjectivity, requiring a system with multiple annotators and a reconciliation mechanism between them (e.g., [10,18]). We believe that without a large budget and/or the ability to involve a community of expert users willing to spend significant time on the task, producing a large ground truth dataset cannot be achieved through a purely manual approach.

[6] `http://piano-e-competition.com/`

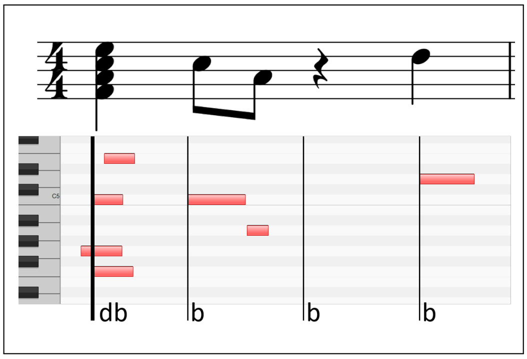

Figure 1. Beat and downbeat annotations produced by our workflow in different cases. The 1st (db) is the median of the onsets of the corresponding notes, the 2nd and the 4th (b) are the onset of the corresponding note, and the 3rd (b) is the mean between the two neighbor annotations.

We therefore propose a workflow (Section 3.1) that allows for the automatic production of annotations from digital musical scores and MIDI performances. The precision of the results is much higher than what would be expected from human-based annotation with a single annotator. Moreover, when the quality of the available scores is high, the workflow does not require any human intervention. Manual intervention is sometimes required to fix problems related to either the digital encoding of the scores or performance errors (Section 3.2).

3.1 Annotation Workflow

The annotation workflow (Figure 3) takes a MIDI performance and a MusicXML score as input, and produces beat, downbeat, key signature change and time signature change annotations for each. Each annotation is aligned to a position (in seconds) in both the MIDI performance and a MIDI score (generated automatically from the MusicXML score), and each downbeat is further aligned with a measure number from the MusicXML score. The workflow is:

1. Expand any repetitions present in the MusicXML score and extract time and key signature changes using music21 [7].
2. Generate the MIDI score using the MuseScore3 MIDI export function.
3. Extract the times of beats, downbeats, and key and time signature changes from the generated MIDI using pretty_midi [33].
4. Align every downbeat from the MIDI score with a measure number in the XML score.
5. Produce a Score2Performance mapping from each note in the MIDI score to each note in the MIDI performance using the algorithm presented in [30].
6. The performance annotations can then be obtained. For each annotation in the MIDI score (from step 3):
    (a) Take the notes with an onset within 20ms of the annotation (there can be multiple notes, e.g. for the downbeat annotation in Figure 1).
    (b) Use the Score2Performance mapping, to obtain the onset times of the corresponding notes in the performance file.
    (c) Compute the median of the onset times of those notes, and use it as the time of the annotation in the performance.
    (d) If no notes are within 20ms of the MIDI score annotation (e.g., in case of a rest), the position is interpolated from neighboring annotations (e.g., the 3rd annotation in Figure 1)

As shown in [13], annotations on rests or multiple non-simultaneous notes (grace-notes, arpeggios) are inherently problematic, even for human annotators. An inspection of our annotations reveals that our workflow generally produces good results. In particular, our use of the median increases robustness while handling non-simultaneous notes.

3.2 Practical issues with erroneous input

While the mapping between the scores and performances in ASAP is such that they have a very good correspondence in general, local problems can still occur in specific cases. These can be caused either by encoding issues in the XML score, or by performance errors in the MIDI.

For our automatic workflow to produce good results, the *content* level of the XML score must be correctly encoded, although we can ignore problems at the *graphical* level (we refer to the model of the multiple levels of information in a musical score proposed in [9]). Many of the problems that we encountered during the automatic creation of ASAP's annotations are tricks used by editors to fix problems at the graphical level at the expense of correctness at the content level: for example, grace notes entered as regular notes with a smaller font, invisible barlines inserted mid-bar, invisible notes or rests, or note heads changed for visual reasons inconsistent with their actual duration.

In some cases, editors are forced to use such tricks because the original score itself does not follow standard notation rules. Figure 2 shows two examples of incorrect measure duration: one from Ravel's *Ondine* (top) and another from Mozart's *Fantasie in C minor* (in 4/4; bottom left). There are many ways to handle such cases. Possibilities include having invisible tuplet markings, having a different time signature than the one displayed, and having an overflowing measure. The latter two techniques create problems for automatic beat extraction. Figure 2 (bottom right) is an example of a key change in the middle of a bar in Beethoven's *Sonata No.29, Op.106*, 2nd movement. One way to encode this situation is to split the bar into 2 separate bars, but this also creates problems for automatic beat extraction in the form of an extra downbeat.

Fortunately, such problems are generally easy to detect since they often result in a measure where the sum of the events does not match the duration defined by the time signature. To be able to produce correct annotations even in the case of these "faulty" measures (around 3% of measures in ASAP's musical scores), we introduce a manual correction step for the MIDI score annotations (Figure 3). This prevents the propagation of such problems down to the performance annotations.

The alignment that we use to generate the Score2Performance mapping [30] is robust against small errors and

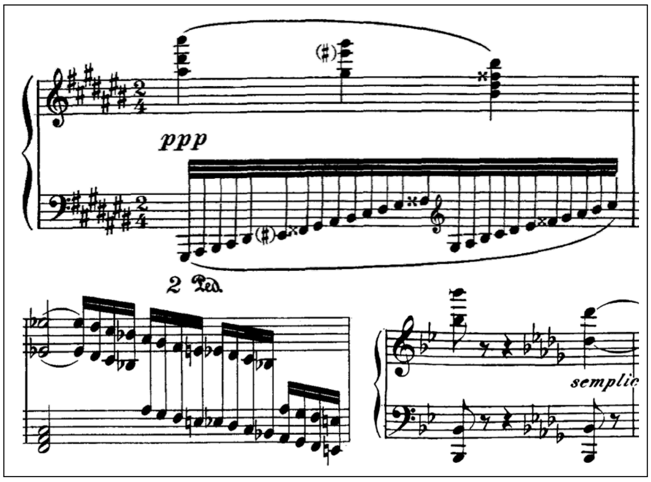

Figure 2. Examples of problems in musical scores, including incorrect bar length (Ravel's *Ondine* (top) and Mozart's *Fantasie in C minor* (in 4/4; bottom left)) and a mid-bar key change (Beethoven's *Sonata No.29, Op.106*, 2nd movement (bottom right)).

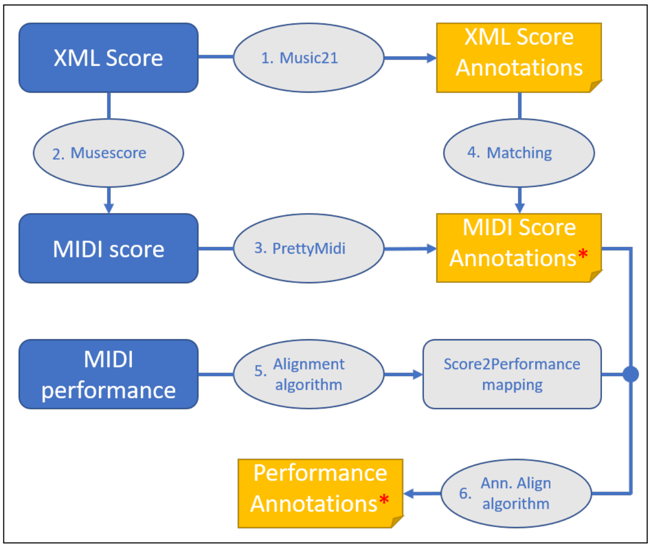

Figure 3. Our annotation workflow; "*" indicates manual correction (see Section 3.2)

note inversions. Nonetheless, small errors in its alignment exist. As the difference between a performance and MIDI score increases, the chance of having an incorrect alignment also increases. This can occur in the case of embellishments (e.g. long trills, or mordents that can be played from the printed note or from the note above) or major performance mistakes. We try to detect these problems automatically by measuring the inter-beat-intervals of each performance and marking the outliers as possible problems. On these outliers (which occurred in around 400 of ASAP's performances), we introduce a final manual correction step. 43 performances contained significant alignment errors that were corrected, and around 2% of annotations had to be moved by less than 1 second.

4. DATASET OVERVIEW

4.1 Dataset content

ASAP contains 222 distinct musical scores and 1068 unique performances of Western classical piano music from 15 different composers (see Table 2 for a breakdown). 548 of the recordings are available as MIDI only, and all the others (520) are provided as MIDI and audio recordings aligned with approximately 3 ms precision. Each score corresponds with at least one performance (and usually more). Every score and performance in ASAP is labeled with metadata including the composer and the title of the piece. We took care to ensure that any two performances of the same piece are labeled with the same title and composer, and no two distinct pieces in ASAP share both.

Each musical score is provided in both MusicXML[7] and MIDI formats. In the MIDI score, the position of all MIDI events are quantized to a metrical grid according to their position in the MusicXML score. Grace notes are represented in MIDI as notes of very short duration. Repetitions in the score are "unfolded" in the MIDI file such that some sections of the MusicXML score may be duplicated in the MIDI score. Except for performance mistakes, there is a one-to-one correspondence between the notes in a MIDI performance and its associated MIDI score.

For each performance and MIDI score, ASAP provides the positions (in seconds) of all beats, downbeats, time signature changes and key signature changes. Time signature changes are annotated only on downbeats. In the case of pickup measures (and pickup measures in the middle of the score) we delay the position of the time signature change annotation to the following downbeat. Similarly, key signature changes are annotated only on beats. Each downbeat is also mapped to a specific measure number in the MusicXML score, which allows for a clear score alignment, even in the case of repetitions.

The dataset and the code used to generate annotations, along with a detailed description of the specific formatting of the dataset and usage examples are available online[8] .

4.2 Origin of the files

The files in ASAP are drawn from multiple sources. The MusicXML scores are from the MuseScore online library[9] , created and uploaded by the users of the MuseScore music notation software. They were first collected and associated to MIDI performances from the Yamaha e-piano competition by the authors of [20]. We manually edited the MusicXML scores using MuseScore3 to correct significant notation errors, and generated quantized MIDI files using MuseScore3's MIDI export utility. The paired audio and MIDI performances come from the MAESTRO dataset [17]. We automatically matched as many of the MAESTRO performances as we could to ones collected by [20], thus associating them with musical scores. The unmatched performances from MAESTRO are not included in ASAP. Finally, we modified 5 of the MIDI performances

[7] https://www.musicxml.com/

[8] https://github.com/fosfrancesco/asap-dataset

[9] https://musescore.com/sheetmusic

Composer	XML/MIDI Score	MIDI Perf.	Audio Perf.
Bach	59	169	152
Balakirev	1	10	3
Beethoven	57	271	120
Brahms	1	1	0
Chopin	34	290	109
Debussy	2	3	3
Glinka	1	2	2
Haydn	11	44	16
Liszt	16	121	48
Mozart	6	16	5
Prokofiev	1	8	0
Rachmaninoff	4	8	4
Ravel	4	22	0
Schubert	13	62	44
Schumann	10	28	7
Scriabin	2	13	7
Total	222	1068	520

Table 2. The composition of ASAP.

by inserting a missing note at the beginning, and 277 more have been cut to obtain more homogeneous pieces (e.g., the Bach Preludes are separated from the Fugues, even though they sometimes come from the same performance).

4.3 Dataset Statistics

ASAP contains performances of classical piano music, a style that can be challenging for beat tracking systems due to the large tempo variations that are often present. Here, we compare the tempo variation of the pieces in ASAP to that of the pieces of datasets commonly used for beat tracking from audio [12,14,18,38]. To quantify the tempo variation for each piece, we first compute the BPM at each beat based on the amount of time between consecutive beats. Then, we compute Δ_{BPM} at each beat as the difference between consecutive BPMs. Finally, we compute the standard deviation of the set of all Δ_{BPM} values in a particular piece, which we call $\sigma(\Delta_{BPM})$. Figure 4 presents the distribution of these standard deviations for each dataset as a Cumulative Distribution Function, which shows the probability that a randomly chosen piece from each dataset has a $\sigma(\Delta_{BPM})$ less than the given value. From the plot, it can be seen that Ballroom and SMC have generally steadier tempos than the other datasets, and that ASAP's steadiest 40% and 50% of pieces roughly match those of Hainsworth and SMC respectively. However, a large portion of the pieces in ASAP have significantly larger tempo variation than any of the compared datasets.

In ASAP, differences can be observed between composers. Even though the pieces in the dataset fall under the broad term "classical music", ASAP is very diverse, containing pieces of various styles written across different periods. This can be observed from differences in average $\sigma(\Delta_{BPM})$ for each composer, as is shown in Figure 5. The values in this plot generally match musicological intuitions about tempo variation for the different composers.

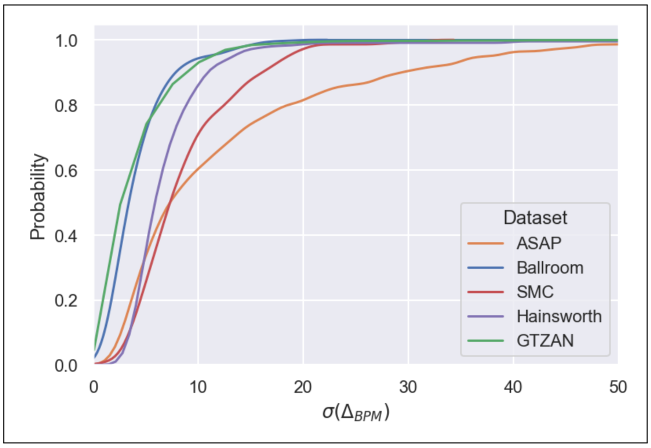

Figure 4. Cumulative Distribution Function of tempo variation $\sigma(\Delta_{BPM})$ of each piece in the compared datasets.

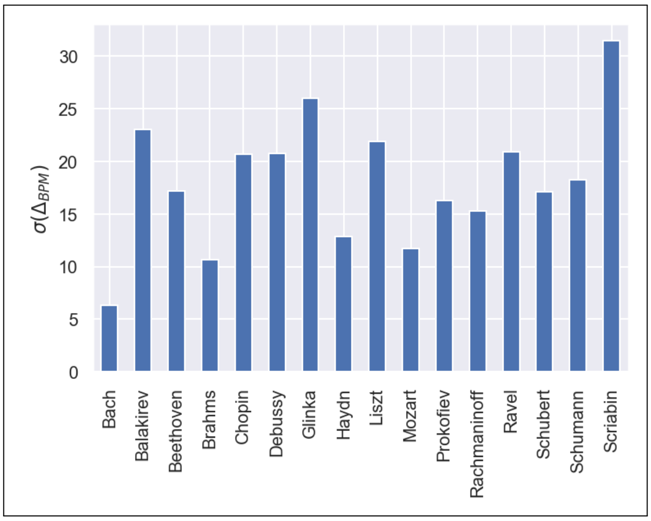

Figure 5. Average tempo variation $\sigma(\Delta_{BPM})$ of the pieces by each composer in ASAP.

5. CONCLUSION

This paper presented ASAP: a new dataset of aligned musical scores and performances of classical piano music. Downbeat, beat, time signature, and key signature annotations are produced using a novel workflow that exploits information present in the musical score to drastically reduce manual annotation time compared to fully manual annotation. ASAP contains over 1000 annotated MIDI performances of classical piano music, over 500 of which are paired with audio from the MAESTRO dataset. To our knowledge, it is the largest dataset of that contains such a fine-grained alignment between scores and performances.

This work has only scratched the surface of what can be done with ASAP. Future work will present further statistical analyses on the data and baseline model performance on tasks for which it can be used: complete AMT and beat tracking as presented, as well as others such as expressive performance analysis and rendering [4]. For complete AMT in particular, the evaluation method is still an open problem, although proposals have been made (e.g., [6,27]).

6. ACKNOWLEDGMENTS

This work was partially funded by ANR France-Quebec MUNIR Project. The first author conducted his research as International Research Fellow of the Japan Society for the Promotion of Science.

7. REFERENCES

[1] Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 36(1):20–30, jan 2019.

[2] Sebastian Böck, Matthew E.P. Davies, and Peter Knees. Multi-task learning of tempo and beat: Learning one to improve the other. In *ISMIR*, 2019.

[3] Jorge Calvo-Zaragoza, Jan Hajic Jr, and Alexander Pacha. Understanding optical music recognition. *Computer Research Repository, abs/1908.03608*, 2019.

[4] Carlos E. Cancino-Chacón, Maarten Grachten, Werner Goebl, and Gerhard Widmer. Computational models of expressive music performance: A comprehensive and critical review. *Frontiers in Digital Humanities*, 5:25, 2018.

[5] Ralf Gunter Correa Carvalho and Paris Smaragdis. Towards end-to-end polyphonic music transcription: Transforming music audio directly to a score. In *WASPAA*, pages 151–155, 2017.

[6] Andrea Cogliati and Zhiyao Duan. A metric for music notation transcription accuracy. In *ISMIR*, pages 407–413, 2017.

[7] Michael Scott Cuthbert, Christopher Ariza, and Lisa Friedland. Feature extraction and machine learning on symbolic music using the music21 toolkit. In *ISMIR*, pages 387–392, 2011.

[8] Valentin Emiya, Roland Badeau, and Bertrand David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1643–1654, aug 2010.

[9] Francesco Foscarin, David Fiala, Florent Jacquemard, Philippe Rigaux, and Virginie Thion. Gioqoso, an online Quality Assessment Tool for Music Notation. In *4th International Conference on Technologies for Music Notation and Representation (TENOR'18)*, Montreal, Canada, May 2018.

[10] Thassilo Gadermaier and Gerhard Widmer. A study of annotation and alignment accuracy for performance comparison in complex orchestral music. *arXiv preprint arXiv:1910.07394*, 2019.

[11] Werner Goebl. Numerisch-klassifikatorische interpretationsanalyse mit dem "Bösendorfer Computerflügel". Master's thesis, Universität Wien, 1999.

[12] Fabien Gouyon, Anssi Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1832–1844, 2006.

[13] Peter Grosche, Meinard Müller, and Craig Stuart Sapp. What makes beat tracking difficult? a case study on chopin mazurkas. In *ISMIR*, pages 649–654, 2010.

[14] Stephen W Hainsworth and Malcolm D Macleod. Particle filtering applied to musical tempo tracking. *EURASIP Journal on Advances in Signal Processing*, 2004(15):927847, 2004.

[15] Mitsuyo Hashida, Toshie Matsui, and Haruhiro Katayose. A new music database describing deviation information of performance expressions. In *ISMIR*, pages 489–494, 2008.

[16] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck. Onsets and frames: Dual-objective piano transcription. In *ISMIR*, 2018.

[17] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*, 2019.

[18] Andre Holzapfel, Matthew EP Davies, José R Zapata, João Lobato Oliveira, and Fabien Gouyon. Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2539–2548, 2012.

[19] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations*, 2018.

[20] Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, and Juhan Nam. VirtuosoNet: A hierarchical RNN-based system for modeling expressive piano performance. In *ISMIR*, pages 908–915, 2019.

[21] Rainer Kelz, Matthias Dorfer, Filip Korzeniowski, Sebastian Böck, Andreas Arzt, and Gerhard Widmer. On the potential of simple framewise approaches to piano transcription. In *ISMIR*, pages 475–481, 2016.

[22] Jong Wook Kim and Juan Pablo Bello. Adversarial learning for improved onsets and frames music transcription. In *ISMIR*, pages 670–677, 2019.

[23] Florian Krebs, Sebastian Böck, and Gerhard Widmer. Rhythmic pattern modeling for beat and downbeat tracking in musical audio. In *ISMIR*, pages 227–232, 2013.

[24] Ugo Marchand and Geoffroy Peeters. Swing ratio estimation. In *Digital Audio Effects (DAFx)*, pages 423–428, 2015.

[25] Andrew McLeod, Eita Nakamura, and Kazuyoshi Yoshii. Improved metrical alignment of MIDI performance based on a repetition-aware online-adapted grammar. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 186–190, 2019.

[26] Andrew McLeod and Mark Steedman. HMM-based voice separation of MIDI performance. *Journal of New Music Research*, 45(1):17–26, 2016.

[27] Andrew McLeod and Mark Steedman. Evaluating automatic polyphonic music transcription. In *ISMIR*, pages 42–49, 2018.

[28] Meinard Müller, Verena Konz, Wolfgang Bogler, and Vlora Arifi-Müller. Saarland music data (SMD). In *Late-Breaking and Demo Session of ISMIR*, 2011.

[29] Eita Nakamura, Emmanouil Benetos, Kazuyoshi Yoshii, and Simon Dixon. Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 101–105. IEEE, 2018.

[30] Eita Nakamura, Kazuyoshi Yoshii, and Haruhiro Katayose. Performance error detection and post-processing for fast and accurate symbolic music alignment. In *ISMIR*, pages 347–353, 2017.

[31] Ryo Nishikimi, Eita Nakamura, Satoru Fukayama, Masataka Goto, and Kazuyoshi Yoshii. Automatic singing transcription based on encoder-decoder recurrent neural networks with a weakly-supervised attention mechanism. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165. IEEE, 2019.

[32] Adriana Olmos, Nicolas Bouillot, Trevor Knight, Nordhal Mabire, Josh Redel, and Jeremy R Cooperstock. A high-fidelity orchestra simulator for individual musicians' practice. *Computer Music Journal*, 36(2):55–73, 2012.

[33] Colin Raffel and Daniel P. W. Ellis. Intuitive analysis, creation and manipulation of midi data with pretty_midi. In *ISMIR Late Breaking and Demo Papers*, 2014.

[34] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

[35] Miguel A. Román, Antonio Pertusa, and Jorge Calvo-Zaragoza. A holistic approach to polyphonic music transcription with neural netwoks. In *ISMIR*, pages 731–737, 2019.

[36] Zhengshan Shi, Craig Stuart Sapp, Kumaran Arul, Jerry McBride, and Julius O. Smith. SUPRA: Digitizing the stanford university piano roll archive. In *ISMIR*, pages 517–523, 2019.

[37] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-U-Net: A multi-scale neural network for end-to-end audio source separation. In *ISMIR*, pages 334–340, 2018.

[38] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

[39] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[40] Christof Weiß, Vlora Arifi-Müller, Thomas Prätzlich, Rainer Kleinertz, and Meinard Müller. Analyzing measure annotations for western classical music recordings. In *ISMIR*, pages 517–523, 2016.

[41] Adrien Ycart and Emmanouil Benetos. A-MAPS: Augmented MAPS dataset with rhythm and key annotations. In *ISMIR Late Breaking and Demo Papers*, 2018.

[42] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.