ACPAS: A DATASET OF ALIGNED CLASSICAL PIANO AUDIO AND SCORES FOR AUDIO-TO-SCORE TRANSCRIPTION

Lele Liu
Queen Mary University of London
lele.liu@qmul.ac.uk

Veronica Morfi
Queen Mary University of London
Sonantic, London
g.v.morfi@qmul.ac.uk

Emmanouil Benetos
Queen Mary University of London
emmanouil.benetos@qmul.ac.uk

ABSTRACT

We create the ACPAS dataset with aligned audio and scores on classical piano music for automatic music audio-to-score transcription research. The dataset contains 497 distinct music scores aligned with 2189 audio performances, 179.8 hours in total. To our knowledge, it is currently the largest dataset for audio-to-score transcription research. We provide aligned performance audio, performance MIDI and MIDI scores, together with beat, key signature, and time signature annotations. The dataset is partly collected from a list of existing automatic music transcription (AMT) datasets, and partly synthesized. Both real recordings and synthetic recordings are included. We provide a train/validation/test split with no piece overlap and in line with splits in other AMT datasets.

1. INTRODUCTION

Automatic music transcription (AMT) is a core research topic in Music Information Retrieval [1, 2]. Given the increasing performance of multi-pitch detection systems, researchers are more frequently working on audio-to-score (A2S) transcription [3–5]. There is, however, a lack of data that limits A2S transcription research. The current largest dataset for A2S transcription research is the ASAP dataset [6], with 222 distinct music pieces, 519 audio performances (from the MAESTRO dataset [7]), and 1067 MIDI performances. Although the dataset provides a large amount of performances, only half of them derive from audio recordings.

Another possible source of aligned music audio and score annotations is the A-MAPS dataset [8], where the audio recordings come from the MAPS dataset [9]. The MAPS dataset is one of the mostly used datasets in AMT research. It provides performance audio recordings and their MIDI ground truth, but without any rhythm or key annotations. The A-MAPS dataset provides an updated version of the MAPS ground truth with beat, key and time signature annotations. However, there are only 266 performances, even less than the audio performances provided in the ASAP dataset.

In this work, we create the ACPAS dataset [1], a dataset with aligned classical piano audio and scores for audio-to-score transcription tasks. The dataset contains 2189 audio performances of 497 distinct music pieces, and is the largest dataset for A2S transcription, to our knowledge. We provide a train/validation/test split where there is no piece overlap between splits, which is also in line with existing train/test splits used or provided in the MAPS and MAESTRO datasets.

2. ACPAS DATASET

2.1 Data Collection

The ACPAS dataset is created based on three sources: 1) the MAPS and A-MAPS dataset; 2) the Classical Piano MIDI [2] (CPM) database; and 3) the ASAP dataset. We first make use of the MAPS recordings and the augmented rhythm and key annotations provided in the A-MAPS dataset. Then, we refer to the CPM database, which is a source database of the MAPS dataset. However, the MAPS dataset only used a part of the CPM pieces. Also, due to the continued update on the CPM database, there are inconsistencies between the MAPS performances and CPM performances even for the same music piece. To enlarge our corpus, we make use of the MIDI files in the CPM database by synthesising the MIDI performances to get the corresponding audio recordings. We also add the ASAP corpus, and further enlarge it by synthesising the MIDI performances without a corresponding audio performance.

Combining data from all the aforementioned sources, we draw a new dataset with 497 distinct music pieces and 2189 audio performances. Due to different sources of data, the performances can be human performances (from the ASAP dataset) or hand-crafted tempos and dynamics to sound like human performances (from MAPS, A-MAPS dataset and CPM database). The performance MIDI, annotations and MIDI scores are collected/modified from the source datasets, and audio performances are partly from the source datasets and partly newly synthesized.

During synthesis, we make use of four different piano models provided in the Native Instrument Kontakt

[1] https://cheriell.github.io/research/ACPAS_dataset
[2] http://www.piano-midi.de

Dataset	Score	Audio Perf.	MIDI Perf.
MAPS	✗	270 (18.1)	270 (18.1)
MAESTRO	✗	1282 (201.2)	1282 (201.2)
A-MAPS	158	266 (17.7)	266 (17.7)
ASAP	222	519 (44.7)	1067 (94.2)
ACPAS	497	2189 (179.8)	2189 (179.8)

Table 1. Comparison on different AMT datasets. Score: number of distinct music pieces. The performance data is provided in the form of *number (duration in hours)*. The MAESTRO dataset here refers to v2.0.0.

Player [3] . We randomly tune the piano fonts to be hard or soft across performances, and add some level of reverberation to simulate real recording conditions. The synthetic audio recordings are rendered to monaural .wav files at 44.1kHz, 16 bit.

2.2 Dataset Content

The dataset is divided into a *Real recording subset* and a *Synthetic subset*:

- The *Real recording subset* covers 578 real recording performances from 1) the two real recording subsets ("ENSTDkCl" and "ENSTDkAm") in the MAPS dataset and corresponding MIDI scores from the A-MAPS dataset, and 2) the performances from the ASAP dataset with audio recordings from the MAESTRO dataset.

- The *Synthetic subset* covers 1611 performances with synthetic audio recordings from the following three sources: 1) performances from the MAPS synthetic subsets and MIDI scores from the A-MAPS dataset; 2) MIDI performances and scores from the ASAP dataset and audio files synthesized from performance MIDIs using Native Instrument Kontakt Player; and 3) MIDI performances and scores from the CPM database and audio files synthesized from performance MIDI using Native Instrument Kontakt Player.

We provide a pre-defined train/validation/test split for the dataset, so there is no piece overlap among splits over the whole dataset. We also take into consideration some train/test splits in existing datasets (MAPS and MAESTRO). The provided split ensures that no test piece in the MAPS and MAESTRO dataset appears in the training or validation split in the ACPAS dataset. This provides the advantage that models trained on ACPAS dataset can be tested using MAPS and MAESTRO test sets. Here, we refer to the test set of MAPS to be the two real recording subsets ("ENSTDkC" and "ENSTDkAm") that are most commonly used for testing, and the test set of MAESTRO to be the test split provided in v2.0.0. After reserving all the test pieces appeared in the MAPS and MAESTRO dataset, we randomly divide the remaining pieces to train/validation.

[3] https://www.native-instruments.com/en/products/komplete/samplers/kontakt-6-player

Subset/Split	Dist. Score	Perf.	Duration (h)
Real recording	215	578	49.0
Synthetic	497	1611	130.8
train	359	1523	127.7
validation	49	184	11.2
test	89	482	40.9
Total	497	2189	179.8

Table 2. ACPAS dataset: statistics across subsets and splits.

During synthesis, we reserve one piano model to synthesize test pieces only. The other three piano models are used across all the three splits.

Individual splits for each music piece and music performance are provided in the dataset metadata. We provide two sets of metadata, a list of all the distinct music pieces and metadata for each performance. The latter covers information including distinct piece ID, composer, title, performance duration, and path to the audio and MIDI files.

2.3 Annotations

For each performance and MIDI score, we provide an annotation file with its beats/downbeats, key signatures and time signatures. The file format is similar to what is provided in the ASAP dataset, with a list of beat times and their corresponding labels including: 1) beat annotation (b for beat, db for downbeat, and bR for an estimated beat where there is e.g. rubato happening); 2) time signature annotation in a string (e.g. '4/4'); and 3) key signature annotation in number of sharps (positive) or flats (negative).

To enable convenient use of the ground truth, we update the MIDI scores collected from the ASAP dataset to fit to the annotated beats, key and time signatures by changing the tick and tempo in the MIDI files. This allows direct use of the MIDI scores as ground truth for some audio-to-score transcription evaluation metrics (e.g. the MV2H metric [10, 11]).

2.4 Dataset statistics

With 2189 classical piano performance audio aligned with 497 distinct music pieces, in total 179.8 hours, the size of the ACPAS dataset surpassed that of other existing A2S datasets. A comparison on the statistics of different AMT datasets can be found in Table 1. Table 2 shows more statistics on the ACPAS dataset across subsets/splits.

3. DISCUSSION

We created the ACPAS dataset by collecting data from several classical piano datasets and synthesising more audio recordings to enlarge the corpus. This new dataset is created to target MIR tasks such as beat/downbeat tracking, key/time signature prediction, and audio-to-score transcription. However, the voice and hand annotations in the score files are not checked and we do not suggest using these as ground truth. In addition, due to the limitation of the scores in MIDI format, the dataset is not suitable for tasks like score formatting.

4. ACKNOWLEDGEMENTS

L. Liu is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by the China Scholarship Council and Queen Mary University of London.

5. REFERENCES

[1] 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.

[2] L. Liu and E. Benetos, "From Audio to Music Notation," in *Handbook of Artificial Intelligence for Music*, E. R. Miranda, Ed. Springer, 2021, pp. 693–714.

[3] K. Shibata, E. Nakamura, and K. Yoshii, "Non-Local Musical Statistics as Guides for Audio-to-Score Piano Transcription," *arXiv preprint arXiv:2008.12710*, 2020. [Online]. Available: http://arxiv.org/abs/2008.12710

[4] M. A. Román, A. Pertusa, and J. Calvo-zaragoza, "Data Representations for Audio-to-Score Monophonic Music Transcription," *Expert Systems With Applications*, vol. 162, p. 113769, 2020. [Online]. Available: https://doi.org/10.1016/j.eswa.2020.113769

[5] L. Liu, V. Morfi, and E. Benetos, "Joint Multi-pitch Detection and Score Transcription for Polyphonic Piano Music," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

[6] F. Foscarin, A. Mcleod, P. Rigaux, F. Jacquemard, and M. Sakai, "ASAP: A Dataset of Aligned Scores and Performances for Piano Transcription," in *ISMIR, International Society for Music Information Retrieval Conference*, 2020.

[7] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," in *ICLR, International Conference on Learning Representations*, 2019. [Online]. Available: http://arxiv.org/abs/1810.12247

[8] A. Ycart and E. Benetos, "A-MAPS: Augmented MAPS Dataset with Rhythm and Key Annotations," in *ISMIR, International Society for Music Information Retrieval Conference, Late-Breaking Demo*, 2018.

[9] 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.

[10] A. Mcleod and M. Steedman, "Evaluating Automatic Polyphonic Music Transcription," in *ISMIR, International Society for Music Information Retrieval Conference*, 2018, pp. 42–49. [Online]. Available: https://www.github.com/apmcleod/MV2H.

[11] A. Mcleod, "Evaluating Non-Aligned Musical Score Transcriptions with MV2H," in *ISMIR, International Society for Music Information Retrieval Conference, Late-Breaking Demo*, 2019.