# ATEPP: A DATASET OF AUTOMATICALLY TRANSCRIBED EXPRESSIVE PIANO PERFORMANCE

**Huan Zhang*** **Jingjing Tang*** **Syed Rifat Mahmud Rafee*** **Simon Dixon** **György Fazekas**
School of Electronic Engineering and Computer Science, Queen Mary University of London
huan.zhang@qmul.ac.uk, jingjing.tang@qmul.ac.uk, s.rafee@qmul.ac.uk

## ABSTRACT

Computational models of expressive piano performance rely on attributes like tempo, timing, dynamics and pedalling. Despite some promising models for performance assessment and performance rendering, results are limited by the scale, breadth and uniformity of existing datasets. In this paper, we present ATEPP, a dataset that contains 1000 hours of performances of standard piano repertoire by 49 world-renowned pianists, organized and aligned by compositions and movements for comparative studies. Scores in MusicXML format are also available for around half of the tracks. We first evaluate and verify the use of transcribed MIDI for representing expressive performance with a listening evaluation that involves recent transcription models. Then, the process of sourcing and curating the dataset is outlined, including composition entity resolution and a pipeline for audio matching and solo filtering. Finally, we conduct baseline experiments for performer identification and performance rendering on our dataset, demonstrating its potential in generalizing expressive features of individual performing style.

## 1. INTRODUCTION

Expressive piano performance has long been explored using data-driven approaches for performance analysis and generation. Recently, more attention has been paid to data-hungry, deep learning techniques, for expressive performance rendering and assessment [1, 2]. Large-scale datasets of expressive piano performances that vary across composition, performers, genres, etc. are demanded by researchers who intend to build comprehensive models and compare different architectures.

Most of the current work that studies expressive piano performances [3–5] uses MIDI rather than audio [6, 7], as MIDI provides easier access to performance attributes including tempo, timing, dynamics, and pedalling. However, datasets that consist of recorded MIDI files from computer-controlled pianos are limited in size and variety. Although promising approaches applied such datasets to train models for rendering human-like piano performances from scores [3,4], researchers were unable to explore performer-specific expressiveness or different schools of playing with deep learning models due to data limitations. Few pay attention to applying deep learning techniques to performer-related tasks such as performer identification [8, 9], style-specific performance rendering [4, 10], and performance style transfer.

Our contribution is three-fold: First, we performed an error analysis for piano performance transcription, comparing state-of-the-art models and verifying the reliability of transcribed performances with listening tests in Section 3. Second, we focus on Western classical piano music and release a dataset with sufficient richness and variety for studying expressiveness and styles across different performers. Our released dataset [1] is a performer-oriented dataset that consists of 11742 virtuoso recordings with 1007 hours of music. Instead of recording MIDI files by computer-controlled piano, we collected our dataset by applying state-of-art piano transcription models such as those by Kong et al. [11] and Hawthorne et al. [12] to transcribe the existing audio recordings of piano performances into MIDI files. More details of our dataset and a reproduction pipeline are presented in Sections 3 and 4. Finally, we demonstrate the application of our dataset to the tasks of performer identification and performance rendering in Section 5. Besides these two tasks, ATEPP can also be utilized in analyzing performance attributes [13–15], comparative study of performances and styles [16], as well as performance visualization [17].

## 2. RELATED WORK

### 2.1 Dataset Requirements for Piano Performance Research

In order to create a comprehensive dataset addressing tasks like assessment and rendering, we discuss the following requirements, with a comparison of existing datasets in Table 1.

- **Multiple performances of the same music:** With the goal of capturing expressive details and common per-

* Equal contributions



[1] Released dataset and supplementary material (Appendix): https://github.com/BetsyTang/ATEPP. The dataset is made available under Creative Commons Attribution 4.0 International Public License (CC BY 4.0).

| Dataset | Size | | | Artist | | Modality | | |
|---|---|---|---|---|---|---|---|---|
| | Performances | Hours | Compositions | Composers | Performers | Perf. MIDI | Score | CER |
| SUPRA [18] | 478 | 52 | 408 | 111 | 153 | ✓ | × | × |
| SMD [19] | 50 | 4.7 | 50 | 11 | unknown | ✓ | × | × |
| MazurkaBL [20] | 2000 | 110 | 44 | 1 | 135 | × | 100% | ✓ |
| Maestro v3.0 [21] | 1276 | 172 | 864 | 60 | 205* | ✓ | × | × |
| CrestMusePEDB [22] | 443 | unknown | 35 | 14 | 12 | ✓ | ✓ | × |
| GP† Curated [23] | 7236 | 875 | 7236 | 1787 | unknown | ✓ | × | × |
| ASAP [24] | 1068 | 92 | 222 | 15 | unknown | ✓ | 100% | × |
| ATEPP | 11742 | 1007 | 1580 | 25 | 49 | ✓ | 43% | ✓ |

**Table 1**. Overview of major symbolic piano datasets. CER: composition entity resolution. *Number obtained from crawling the Piano-e-Competition website for performer names and aligning with Maestro data. †GP stands for GiantMIDI-Piano.

formance idioms, comparative study of performance requires multiple versions of the same piece of music, ideally by multiple performers. In the past, datasets with very limited numbers of pieces were recorded and organized by researchers, such as the Mozart [25, 26], Schumann Träumerei [27] and Schubert [8] datasets. The non-trivial task of Composition Entity Resolution **(CER)**, involving the process of automatically aligning the complex naming schemes of classical music, is the major challenge of obtaining multiple performances of the same music at a larger scale. We will detail our CER process in Section 4.1. Among the existing datasets, only the CHARM Mazurka dataset [2] offers CER.

- **Representation:** While the audio recording most faithfully documents a performance, complex processing is needed to extract the expressive attributes from the waveform [28]. MazurkaBL [20] contains many precalculated features that are provided for the Mazurka dataset. Meanwhile, MIDI can serve as a mid-level, piano-roll like representation of piano performance actions. The SUPRA [18] dataset contains expressive MIDI digitised from pneumatic piano rolls, while SMD [19], Maestro [21] and CrestMusePEDB [22] all contain MIDI recorded from Yamaha Disklaviers.
- **Repertoire and diversity:** Given that piano performance traditions are largely associated with the Western classical music paradigm, SMD [19], Maestro [21] and CrestMusePEDB [22] all include standard repertoire from Baroque to Late-Romantic era, while GiantMidi-Piano [23] includes non-standard pieces that span 1.7k composers. The CHARM Mazurka dataset is a great example allowing for multiple-performance comparison, however its repertoire consists only of 49 mazurkas by Chopin.
- **Symbolic score:** A high-level representation of the composition score is typically needed in tasks such as performance rendering [3]. Expressive deviations can be observed by comparing with the quantized, dead-pan score. MazurkaBL [20] and ASAP [24] contain symbolic scores in MusicXML format.
- **Size:** Large datasets are essential for training deep neural networks. Among existing datasets, only GiantMidi-Piano [23] has more than 200 hours of piano music.

[2] `http://www.charm.rhul.ac.uk/index.html`, accessed 12 May 2022.

### 2.2 Automatic Piano Transcription

Empowered by deep learning models, recent automatic piano transcription systems can aid expressive performance research by outputting precise measurements of dynamics, timing and tempo at the note-level. The Onsets and Frames transcription model [29] combined framewise pitch detection with onset detection, to produce a full piano roll with velocity. The High-Resolution model [11] improved precision by regressing the exact timestamp of each note. A recently proposed generic encoder-decoder architecture [12] that exploits language-like modeling achieved model simplicity while retaining performance.

## 3. TRANSCRIPTION AND POST-PROCESSING FOR EXPRESSIVE PERFORMANCE

### 3.1 Common Errors Introduced by Transcription

We categorize common transcription errors into the following three rough categories: harmonic error, segmented note, and mis-touched short note. These errors are obtained by transcribing the performances from the Mazurka dataset using the High-Resolution model [11], and then aligning with its symbolic score in MusicXML using the algorithm by Nakamura et al. [30]. This algorithm aligns two signals, reference and performance, using hidden Markov models (HMMs), detects performance errors from the first alignment result, and then employs a merged-output HMM [31] to correct the errors.

1. **HE**: Harmonic errors (fifths and octaves): The most common type of transcription error is falsely detecting or failing to detect notes that are harmonically related to other played or detected notes. Usually these are missing or extra octaves or fifths, and they result from the overlap of the harmonic series of the pitches.
2. **SN**: Segmented notes: One continuous note being transcribed into two segments with a small (<10ms) gap between offset and onset. This error might come from amplitude modulation [32].

| Model | HE | SN | MS | Other |
|---|---|---|---|---|
| High-Resolution [11] | 3.2% | 1.5% | 1.2% | 5.6% |
| OnsetsFrames [29] | 4.2% | 2.4% | 0.1% | 6.7% |
| Seq2Seq [12] | 8.1% | 2.9% | 0.3% | 7.9% |

**Table 2**. Transcription note error rate (aligned with symbolic score) on the Mazurka dataset.

3. **MS**: Mis-touched short notes: The spurious, short notes (<16ms) that appear randomly in transcription.

In Table 2, we quantitatively evaluate the presence of these error types on the Mazurka dataset. Given that no performance ground truth exists for this data, we rely on the MusicXML score and the assumption that the performances match the same score version. Among the three most recent deep-learning based transcription models [11, 12, 29], the High-Resolution model [11] makes the fewest errors overall, but generates more short notes compared to the other models.

Besides notewise errors, another factor of concern for transcribed MIDI is the note duration. From the inference output [11], the notes' durations are elongated to achieve a sustain effect that is usually implemented by sustain pedal in reality. Whilst such elongation doesn't make an audible difference in MIDI rendering software, accurate end positions of each note are required in piano performance analysis.

### 3.2 Joint Note-Pedal Training

With the goal of reconciling the sustain effect from both pedal and keys, as well as achieving more accurate note offsets, we modify the original High-Resolution model [11] with joint note-pedal training. As the first piano transcription model that incorporated the sustain pedal into training, the High-Resolution model trained separate networks for key activity and sustain pedal (with binary velocity), while extending the note offset to match the pedal off timestamp. In joint note-pedal training, 88 keys and 3 pedal channels are combined to output 91 prediction classes with velocity for each channel, and the extension of note offsets is removed. In this case, during training the sustained effect would be conditioned on both key-down duration as well as pedal controls.

As shown in Figure 1, what we model is the key action from the pianist instead of the string damping time of the note (that can either come from the sustain pedal or key action), which deviates from the traditional transcription task. With other training parameters unchanged, the onset F1 score ($tol = 50\,ms$) achieved is 92.1% after 300k iterations, and onsets and offsets evaluation achieved 68.2%. Note that the evaluation results are much lower than the original results, as we are attempting to learn patterns of behaviour that are not present in the audio.

### 3.3 Score-Alignment and Correction

As described in Section 3.1, a score-performance alignment algorithm [30] is employed to automatically find

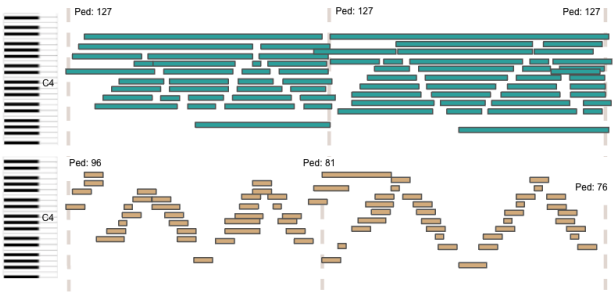


**Figure 1**. Output pianoroll comparison of the original High-Resolution model (top) and joint note-pedal version (bottom). Dashed lines represent pedal-on messages with velocity.

transcription errors with reference to a score. As a post-processing step, we correct the differences according to the alignment. Extra notes (those in transcription but not in score) are deleted, mismatching notes (aligned with pitch error) are corrected to the pitch given in the score, and missing notes (those in score but not in transcription) are interpolated and written back to MIDI according to the following rule:

$$g(i) = g(i - \Delta_p) + (g(i + \Delta_n) - g(i - \Delta_p))\frac{\Delta_p}{\Delta_p + \Delta_n}, \quad (1)$$

where $g(i)$ is the onset or offset timestamp of the missing note at beat $i$, and $\Delta_p, \Delta_n$ represent the beat distances between the missing note and the previous or next existing notes, respectively.

### 3.4 Listening Evaluation

In order to evaluate the perceptual quality of transcribed piano music, we perform a subjective listening test where participants rate the similarity of reproduced MIDI and the reference recording. In the test, we compare the ground truth with 4 transcribed MIDI renderings: the original High-Resolution transcription system [11] (**C1**), the joint note-pedal model described in Section 3.2 (**C2**), a score-corrected version of C2 as described in Section 3.3 (**C3**), and the language-model transcription system [12] (**C4**). All MIDI performances are rendered using a KAWAI CA49 electric piano and recorded using Zoom H4n Pro Recorder. The recordings are then processed with basic noise-reduction in Audacity.

Participants in the listening test are presented with five *20 s* classical piano excerpts with varying style (Q1-Liszt, Q2-Debussy, Q3-Bach, Q4-Rachmaninov, Q5-Mozart; see appendix for music passages). The test is conducted using the MUSHRA protocol [33], each with 5 recordings (reference plus 4 stimuli). Participants are asked which transcribed stimulus sounded closer to the reference on a 100-point scale. During the test, we explicitly ask participants to ignore the timbral or acoustic differences but make judgements based on the expressive differences between the stimuli such as dynamics and timing.

We collected 1075 ratings from 43 listeners. Half of

| | Q1 | Q2 | Q3 | Q4 | Q5 | Overall |
|---|---|---|---|---|---|---|
| Reference | 4.42±0.24 | 4.17±0.29 | 4.24±0.31 | 4.28±0.31 | 4.46±0.27 | 4.30±0.12 |
| C1 | **4.12±0.38** | 3.52±0.37 | 3.88±0.32 | 3.60±0.41 | 3.88±0.4 | 3.81±0.16 |
| C2 | 3.83±0.42 | **3.86±0.39** | **4.28±0.31** | **3.97±0.42** | **4.06±0.45** | **4.01±0.17** |
| C3 | 3.44±0.37 | 2.96±0.36 | 3.44±0.35 | 3.32 ±0.42 | 3.76±0.41 | 3.38±0.17 |
| C4 | 3.88±0.34 | 2.32±0.47 | 3.60±0.29 | 3.84±0.33 | 3.68±0.37 | 3.46±0.18 |

**Table 3**. Results of listening test. The mean opinion scores (MOS) and 95% confidence intervals are reported.

our listeners reported over 5 years of piano performing experience. Table 3 shows the mean opinion scores (converted to a 5-point scale) from the ratings. According to the Wilcoxon signed-rank test ($p < 0.05$), all stimulus groups differ significantly from the reference. Among the stimuli, C2 is preferred by the listeners, while C3 and C4 have significantly lower ratings compared to the other two groups of stimuli. In free text responses, the score-corrected transcription received negative comments such as *unnatural* and *abrupt*. Consequently, we use the C2, note-pedal jointly trained model for transcribing our dataset. The result also shows a perceptual difference of transcription quality across music styles, demonstrating the bias of transcription models: transcribed fast, arpeggio-heavy passages are rated with lower perceptual quality. But for slow, sparse textures, good transcriptions sound much closer to the reference.

## 4. DATASET OVERVIEW

### 4.1 Data Collection and Curation

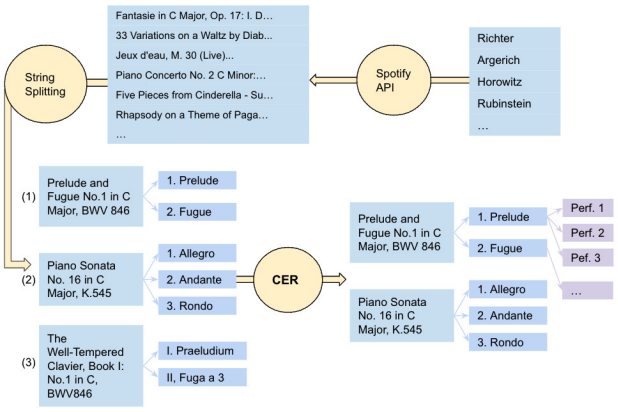


**Figure 2**. Data curation pipeline.

Our data collection pipeline is presented in Figure 2. Starting with 49 world-renowned pianists, metadata (including composer, performer, album, title and track duration) of their discography was obtained using the Spotify API [3]. After filtering out non-solo keywords such as *concerto* or *trio*, a composition-movement hierarchy was built.

As discussed in Section 2.1, the next challenging step was achieving *Composition Entity Resolution (CER)*, defined as finding out which tracks correspond to the same piece of music, given the variety of naming conventions in classical music. For example, compositions (1) and (3) in Figure 2 actually correspond to the same work, while their title differs so much that a simple string similarity match is not sufficient to resolve them as identical.

We address CER using three steps: **1)** Language-specific mapping. We manually compile a dictionary for interchangeable terms, such as *Prelude* ↔ *Praeludium*. **2)** Unique identifier extraction. Unique information such as key and catalogue number (*Opus*, *BWV*, *K.*, *D.*, etc.) are extracted from the title string. **3)** Fuzzy string matching. For both composition title and movement title, we use normalized Levenshtein distance [34] to compute similarity scores. Note that such string matching is not always reliable, as generic names like *Piano Sonata* are extremely frequent in our discography. Combining the three steps, our composition entity resolution is described in Algorithm 1, where inputs include composition title $C$ and movement title $M$ as well as duration $D$. Based on the organized metadata, we download each track from a corresponding open source audio at YouTube Music, while the online metadata is again validated by the same CER algorithm.

**Algorithm 1** Composition Entity Resolution

```
# UniqueInfo extracts canonical key and composer-specific catalogue number.
for k1, k2 in UniqueInfo(C1, C2) do
    if k1 ≠ k2 then return False
    end if
end for
S_c ← 1 − (Levenshtein(C1, C2))/ max(|C1|, |C2|)
S_m ← 1 − (Levenshtein(M1, M2))/ max(|M1|, |M2|)
S_d ← abs(D1 − D2) / max(D1, D2)
S ← (S_c + S_m)/2 − S_d
return S ≥ 0.6
```

### 4.2 Audio Matching by Chroma Features

Besides metadata linking using CER, we also match tracks by downloaded audio content to ensure the same piece of music is being performed. Within each group of performances, we apply Chen et al.'s cover song detection algorithm [35] to compare each performance with a reference. We first extract the Harmonic Pitch Class Profile (HPCP) [36] from the reference and the target performance

[3] https://developer.spotify.com/documentation/web-api/

audio. Next, we use the $Q_{max}$ measure, which represents the maximum value of a cumulative matrix computed from the HPCP descriptors of two performances [37]. The similarity between two performances is defined by Eq. 2. We only retain performances whose similarity is larger than 0.9 to build the ATEPP dataset.

$$Sim = 1 - Q_{max}, \quad 0 < Q_{max} < 1 \tag{2}$$

### 4.3 Applause Filtering with CNN

We also need to filter out any sound that might not be part of a solo piano performance. The most prominent ones are applause and speech from live recordings, which would be transcribed as random pitch. We train a deep learning model based on the Musicnn [38] architecture to filter out any non-solo-piano segment. For each 1 $s$ segment, the probability of non-piano sound is predicted using a binary classifier. In training, a subset of AudioSet [39] with various environmental sounds is used as negative examples, and solo piano recordings are used as positive examples. With a binary tag-gram inferred from the audio, our post-processing step searches for the timestamp of the longest continuous non-solo segment at the beginning and the end to remove from the audio file.

Among the 11742 tracks, 567 of them are detected with starting or ending applause, and were subsequently cleaned. This was followed by a manual verification process by listening to ensure the audio split was accurate.

### 4.4 MusicXML Score

Given that the musical score is also important for music research, we collect scores in MusicXML format that correspond to our performance data. 228 files are drawn from the ASAP dataset [24], and 90 files from the MuseScore [4] online library, crowd-sourced by the users of MuseScore software. This results in a total of 319 movements, corresponding to 5124 tracks in our dataset (43% of all tracks). The score-performance correspondence is first determined automatically by name matching followed by manual correction.

### 4.5 Content and Statistics

The ATEPP dataset contains 11742 tracks of 1580 movements. The tracks overlap only 0.2% of the GiantMidi-Piano dataset [23]. Figure 3 shows a breakdown of the movement-performance distribution. Among 1580 movements, 44% have more than 5 performances, providing us with rich data for studying different interpretations of the same piece of music.

In addition, we show a distribution of the top 25 pianists in our dataset in Figure 4, where Sviatoslav Richter contributes the most data in ATEPP. Composer-wise, solo piano works from 25 Western classical composers are included in our dataset, ranging from the Baroque to the Modern era (a full composer breakdown is given in the appendix).

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

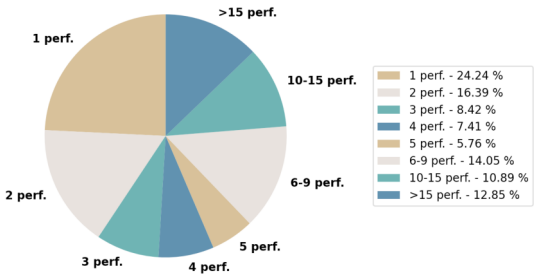


**Figure 3**. Distribution of movements by number of performances. E.g. 12% of our data have more than 15 performances.

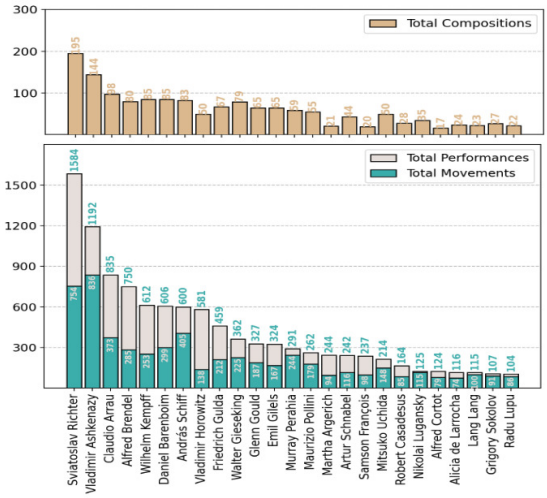


**Figure 4**. Distribution of the top 25 pianists' performances in the ATEPP dataset.

## 5. DATASET APPLICATIONS

### 5.1 Performer Identification

Distinguishing virtuoso performers using computational models has long been studied by researchers who focus on expressive parameters of music performances. Data-driven approaches such as traditional machine learning methods and feature distribution comparison have been applied to this task [8, 9, 40]. However, none of the existing studies have applied deep learning methods to performer identification, due to the lack of large-scale datasets with overlapping performances by different performers.

With the ATEPP dataset, we are now able to train deep neural networks to identify different performers. We choose four subsets from the ATEPP dataset, only considering performers with over 100 performances. For the Mixture subset, we only consider compositions that have more than 15 different performances. We also create three composer-specific subsets (L. Beethoven, F. Chopin, and J. S. Bach) to remove the bias of performer-composer correlation. The train-validate-test split is 8:1:1.

We extracted note-level features including onset and offset time, velocity, and pitch number from the MIDI files without any expression-related preprocessing. We stacked the sequences of those features and input them into a 1D

| Subset | Pianists | Size | Acc. | F1-score |
|---|---|---|---|---|
| Mixture | 16 | 4676 | 0.47 | 0.45 |
| Beethoven | 12 | 3078 | 0.48 | 0.46 |
| Chopin | 5 | 973 | 0.55 | 0.54 |
| Bach | 5 | 1019 | 0.59 | 0.55 |

**Table 4**. Performer identification results.

convolution neural network (see Appendix for the network details). The model was trained on four subsets at a learning rate of $10^{-4}$ with a decay weight of $10^{-8}$.

With all subsets achieving over 0.45 F1 score in Table 4, our baseline model demonstrates the capability of learning individual performing style if given enough data, as well as generalizing across different compositions.

Moreover, the results from composer-specific datasets show that we can achieve comparable results even when performer-style correlation (e.g. Horowitz's repertoire concentrates on Romantic styles while Gould almost exclusively plays Bach) was removed. Confusion matrices are provided in the appendix as well as precision, recall, and F1-score for each performer.

### 5.2 Performance Rendering with VirtuosoNet

There has been a growing research interest in quantifying and modelling expressive performance using computational models [1], including understanding of how humans perform [41], automatically generating expressive performances [42], and rendering expressive performances from a score [3]. Again, ATEPP contributes to such tasks by providing expressive performance data on a large scale.

In this study, we show the utility of our dataset using the performance rendering model VirtuosoNet [3]. The model consists of an RNN with a hierarchical attention network to model note, beat and measure level hierarchy of music and a conditional variational autoencoder (CVAE) to model the expressive performance. We selected two pianists' Beethoven performances with over 300 tracks from the ATEPP dataset, and altered the model slightly to render a performance in the style of a particular performer by concatenating a performer label vector to the latent vector. During training, we use the same hyper-parameter setting as the original paper.

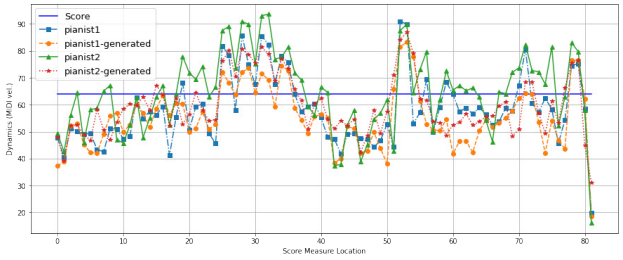


**Figure 5**. Measure level dynamic variation of Beethoven's Piano sonata No. 3 in C, Op. 2 No. 3, Mvt. II, as performed and generated in the styles of pianists 1 and 2. The flat line depicts the default dynamic level provided in the score file.

The trained model takes a series of note-level score features extracted from MusicXML and a performer name as input, and predicts the corresponding note-level performance features of that performer. Figure 5 shows the real and predicted (pianist1-generated & pianist2-generated) performances of an out-of-sample music, Beethoven's Piano Sonata No. 3 in C Major, Op. 2 No. 3, Mvt. II, by pianist 1 and 2 in terms of dynamics. The flat line represents the default dynamic level of a non-expressive, mechanical rendition of the score.

We can observe that both actual performances and generated performances (in the style of pianist1 and pianist2) tend to deviate from the default interpretation in a similar way; this can be a demonstration of common performance practice. For individual interpretations, we calculated the cross-correlation of the dynamic variation between the actual performance and the generated version using Pearson's correlation coefficient (r). We found that both actual performances from pianist1 and pianist2 are highly correlated ($r > 0.75$) with their generated counterparts, respectively. In addition, we computed the correlation coefficient between the actual performance of pianist1 and the generated performance of pianist2 and vice-versa. Both of them provide a lower correlation coefficient ($r<0.6$). This demonstrates that deep learning architectures are capable of learning some of the expressive techniques of each individual pianist from respective training data.

## 6. CONCLUSION AND FUTURE WORK

This paper presents ATEPP, a large-scale dataset of 11,742 expressive piano performances by 49 virtuoso pianists. Nearly half of the compositions are provided with scores in MusicXML format. All of the performances were transcribed by a piano transcription model which was trained jointly with pedals and keys. We performed a listening test to evaluate the reliability of the transcription algorithm. To our knowledge, our dataset is the largest dataset of expressive piano performance MIDI with robust metadata of classical music, derived via Composition Entity Resolution (CER). We presented our baseline experiments for performer identification and performer-oriented expressive performance rendering. The results demonstrate that the ATEPP dataset enables us to study expressive features of individual performing styles with deep learning methods.

In the future, we will consider a hierarchical database system that comprehensively links performances with performer, composer, and composition entities through an automatic CER process. A more balanced dataset is planned as well as an extended version with more variety across the skill level of performers.

## 7. ACKNOWLEDGMENTS

This work is supported by the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, funded by UK Research and Innovation [grant number EP/S022694/1]. Jingjing is also funded by China Scholarship Council (CSC) [grant number 202008440382].

## 8. REFERENCES


[1] C. E. Cancino-Chacón, M. Grachten, W. Goebl, and G. Widmer, "Computational models of expressive music performance: A comprehensive and critical review," *Frontiers in Digital Humanities*, vol. 5, no. October, pp. 1–23, 2018.

[2] A. Lerch, C. Arthur, A. Pati, and S. Gururani, "Music performance analysis: A survey," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[3] D. Jeong, T. Kwon, Y. Kim, K. Lee, and J. Nam, "Virtuosonet: A hierarchical RNN-based system for modeling expressive piano performance," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 908–915.

[4] A. Maezawa, K. Yamamoto, and T. Fujishima, "Rendering music performance with interpretation variations using conditional variational RNN," *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 855–861, 2019.

[5] P. Ramoneda, M. Miron, and X. Serra, "Piano fingering with reinforcement learning," 2021.

[6] H.-W. Dong, C. Zhou, T. Berg-Kirkpatrick, and J. Mcauley, "Deep performer: Score-to-audio music performance synthesis," in *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[7] Y. Wu, E. Manilow, Y. Deng, R. Swavely, K. Kastner, T. Cooijmans, A. Courville, C.-Z. A. Huang, and J. Engel, "MIDI-DDSP: Detailed control of musical performance via hierarchical modeling," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=UseMOjWENv

[8] S. R. M. Rafee, G. Fazekas, and G. A. Wiggins, "Performer identification from symbolic representation of music using statistical models," in *Proceedings of the International Computer Music Conference (ICMC)*, 2021.

[9] E. Stamatatos and G. Widmer, "Automatic identification of music performers with learning ensembles," *Artificial Intelligence*, vol. 165, pp. 37–56, 2005.

[10] A. Tobudic and G. Widmer, "Learning to play like the great pianists," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2005, pp. 871–876.

[11] 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 Transactions on Audio Speech and Language Processing*, vol. 29, pp. 3707–3717, 2021.

[12] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. Engel, "Sequence-to-sequence piano transcription with transformers," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[13] M. Bernays and C. Traube, "Expressive production of piano timbre: Touch and playing techniques for timbre control in piano performance," *Proceedings of the Sound and Music Computing Conference*, no. August 2013, pp. 341–346, 2013.

[14] A. Robertson, "Decoding tempo and timing variations in music recordings from beat annotations," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, 2012.

[15] B. H. Repp, "Acoustics, perception, and production of legato articulation on a digital piano," *The Journal of the Acoustical Society of America*, vol. 97, 1995.

[16] C. S. Sapp, "Comparative analysis of multiple performances," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2007.

[17] M. Grachten, W. Goebl, S. Flossmann, and G. Widmer, "Phase-plane representation and visualization of gestural structure in expressive timing," *Journal of New Music Research*, vol. 38, no. 2, pp. 183–195, Jun. 2009.

[18] Z. Shi, C. S. Sapp, K. Arul, J. McBride, and J. O. Smith, "SUPRA: Digitaizing the stanford university piano roll archive," in *Proceeding of the 20th International Society on Music Information Retrieval (ISMIR)*, 2019.

[19] V. Konz, W. Bogler, and V. Arifi-M, "Saarland music data," *Late-Breaking and Demo Session of the International Conference on Music Information Retrieval (ISMIR)*, 2011.

[20] K. Kosta, O. F. Bandtlow, and E. Chew, "MazurkaBL: Score-aligned loudness, beat, and expressive markings data for 2000 chopin mazurka recordings," in *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'18*, 2018, pp. 85–94.

[21] 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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019, pp. 1–12.

[22] M. Hashida, E. Nakamura, and H. Katayose, "CrestMusePEDB 2nd edition: Music performance database with phrase information," in *Proceedings of the 15th Sound and Music Computing (SMC) Conference*, 2018.

[23] Q. Kong, B. Li, J. Chen, and Y. Wang, "GiantMIDI-Piano : A large-scale midi dataset for classical piano music," *Transactions of the International Society for*

*Music Information Retrieval*, vol. 5, no. 1, pp. 87–98, 2022.

[24] 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)*, 2020.

[25] W. Goebl and S. Dixon, "Analysis of tempo classes in performances of mozart sonatas," in *Proceedings of the VII International Symposium on Systematic and Comparative Musicology and III International Conference on Cognitive Musicology*, 2001, pp. 65–76.

[26] A. Tobudic and G. Widmer, "Relational IBL in classical music," *Machine Learning*, vol. 64, no. 1-3, pp. 5–24, 2006.

[27] B. H. Repp, "The dynamics of expressive piano performance: Schumann's "Traumerei" revisited," *The Journal of the Acoustical Society of America*, pp. 641–50, 1996.

[28] G. Widmer, S. Dixon, W. Goebl, E. Pampalk, and A. Tobudic, "In search of the Horowitz factor," *AI Magazine*, vol. 24, no. 3, pp. 111–130, 2003.

[29] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 50–57, 2018.

[30] E. Nakamura, K. Yoshii, 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)*, 2017.

[31] E. Nakamura, N. Ono, Y. Saito, and S. Sagayama, "Merged-output hidden markov model for score following of midi performance with ornaments, desynchronized voices, repeats and skips," in *Proceedings of the 11st Sound and Music Computing Conference 2017, SMC*, 2017.

[32] T. Cheng, S. Dixon, and M. Mauch, "Modelling the decay of piano sounds," in *Proceedings of International Conference on Acoustic, Speech and Signal Processsing (ICASSP)*. IEEE, 2015.

[33] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[34] W. Heeringa, "Measuring dialect pronunciation differences using levenshtein distance," Ph.D. dissertation, University of Groningen, 2005.

[35] N. Chen, W. Li, and H. Xiao, "Fusing similarity functions for cover song identification," *Multimedia Tools and Applications*, vol. 77, no. 2, pp. 2629–2652, 2018.

[36] E. Gomez, "Tonal description of music audio signals," Ph.D. dissertation, University of Pompeu Fabra, 2006.

[37] J. Serrà, X. Serra, and R. G. Andrzejak, "Cross recurrence quantification for cover song identification," *New Journal of Physics*, vol. 11, 2009.

[38] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 637–644, 2018.

[39] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. Channing Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[40] R. Ramirez, E. Maestre, and X. Serra, "Automatic performer identification in commercial monophonic jazz performances," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1514–1523, 2010.

[41] C. E. Cancino-Chacón, T. Gadermaier, G. Widmer, and M. Grachten, "An evaluation of linear and non-linear models of expressive dynamics in classical piano and symphonic music," *Machine Learning*, vol. 106, no. 6, pp. 887–909, 2017.

[42] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan, "This time with feeling: learning expressive musical performance," in *Neural Computing and Applications*, vol. 32, no. 4, 2018, pp. 955–967.