PIAST: A Multimodal Piano Dataset with Audio, Symbolic and Text

Hayeon Bang[1] Eunjin Choi[1] Megan Finch[1] Seungheon Doh[1]
Seolhee Lee[2] Gyeong-Hoon Lee[2] Juhan Nam[1]

[1]Graduate School of Culture Technology, KAIST, South Korea
[2]NCSOFT, South Korea
{hayeonbang, jech, meganelisabethfinch, seungheondoh, juhan.nam}@kaist.ac.kr, {seolhee, ghlee0304}@ncsoft.com

Abstract

While piano music has become a significant area of study in Music Information Retrieval (MIR), there is a notable lack of datasets for piano solo music with text labels. To address this gap, we present PIAST (PIano dataset with Audio, Symbolic, and Text), a piano music dataset. Utilizing a piano-specific taxonomy of semantic tags, we collected 9,673 tracks from YouTube and added human annotations for 2,023 tracks by music experts, resulting in two subsets: PIAST-YT and PIAST-AT. Both include audio, text, tag annotations, and transcribed MIDI utilizing state-of-the-art piano transcription and beat tracking models. Among many possible tasks with the multimodal dataset, we conduct music tagging and retrieval using both audio and MIDI data and report baseline performances to demonstrate its potential as a valuable resource for MIR research.

1 Introduction

Piano music presents unique opportunities for music research due to its ability to express diverse styles using a single instrument and its superior transcription performance. Given these characteristics, it has become a significant area of study in Music Information Retrieval (MIR), encompassing tasks such as classification (Hung et al., 2021; Chou et al., 2021), and music generation with various conditions (Wu and Yang, 2023; Choi and Lee, 2023). While these tasks require datasets that combine piano audio with various modalities such as MIDI, sheet music, or text, there is a notable scarcity of such comprehensive multimodal piano datasets.

However, existing multimodal music datasets, particularly music-text datasets, rarely focus exclusively on piano music, and piano solo pieces comprise only a small portion of general music-text datasets. For instance, in the ECALS Dataset (Doh et al., 2023), a subset of the Million Song Dataset (Bertin-Mahieux et al., 2011), the number of piano solo tracks is very limited. We observed that excluding tracks tagged with instruments other than the piano or genres that could not be solely represented by the piano, only approximately 0.46% of the entire dataset can be identified as piano solo music.

Several piano datasets, such as MAESTRO (Hawthorne et al., 2019), have been developed in recent years, which provide classical piano performances primarily used for piano transcription. Another classical piano dataset, GiantMIDI (Kong et al., 2022), is also commonly used in transcription tasks. Other datasets like Pop1K7 (Hsiao et al., 2021) focus on the performance generation of pop piano music, while PiJAMA (Edwards et al., 2023) is employed for performer identification tasks with their jazz piano data. However, these datasets are confined to a single genre and lack text labels. This absence of genre diversity within a single dataset and the lack of textual information underscores the need for a piano dataset with text information.

Some piano datasets contain emotion labels, such as EMOPIA (Hung et al., 2021) and VGMIDI (Ferreira and Whitehead, 2019). However, these datasets are annotated only with emotion information based on either Russell's four quadrants (Hung et al., 2021) or the valence-arousal model (Ferreira and Whitehead, 2019). This limited annotation approach lacks the rich textual descriptions needed for text-based MIR tasks.

To address the limitations, we present multimodal piano music data with rich text annotations and transcribed MIDI. To build the dataset, we first created a piano-specific taxonomy with 31 tags that include genre, emotion, mood, and style information to encompass the broad and diverse musical range that the piano can express. Based on this taxonomy, we collected data from YouTube, transcribed it to MIDI format, and conducted an annotation process.

The PIAST dataset consists of two subsets: **PIAST-YT**, 9,673 tracks collected from YouTube, providing audio and text information (titles, tags, and descriptions), and **PIAST-AT**, 2,023 tracks with annotations by music experts. This dual approach ensures both breadth and accuracy in the dataset. Additionally, PIAST includes transcribed performance MIDI data alongside audio and text, enhancing its capabilities beyond existing methods (Hsiao et al., 2021).

This paper details the dataset collection process and analyzes the data. We present baseline results for piano music annotation and retrieval tasks, utilizing the PIAST-YT and the PIAST-AT datasets across audio and MIDI domains. The PIAST dataset is available in our online repository[1], and the source code for the experiments can be found on GitHub[2].

2 Dataset

2.1 Taxonomy for Piano Music

To encompass and precisely define the range of expressions possible in solo piano music, we constructed a comprehensive taxonomy considering genre, emotion, mood, and style tags. We classified genres suitable for solo piano music into four categories: jazz, classical, new-age, and pop piano covers, defining sub-genres within each. The detailed classification of the classical genre was not included in this dataset due to the extensive range and complexity unique to classical music. For emotion and mood taxonomy, we combined vocabularies from four existing music datasets with emotion tags (Turnbull et al., 2007; Rouhi et al., 2019; Aljanaki et al., 2017; Choi et al., 2022)), eliminating overlaps. Seven music experts who majored in music composition rated the tags on a 1-5 Likert scale for their suitability in describing solo piano music. We included only words scoring 3.5 or higher and established a taxonomy of 39 tags. After the first annotation process, we removed tags with excessively high co-occurrence or low selection frequencies, resulting in a final specialized taxonomy of 31 words for piano music.

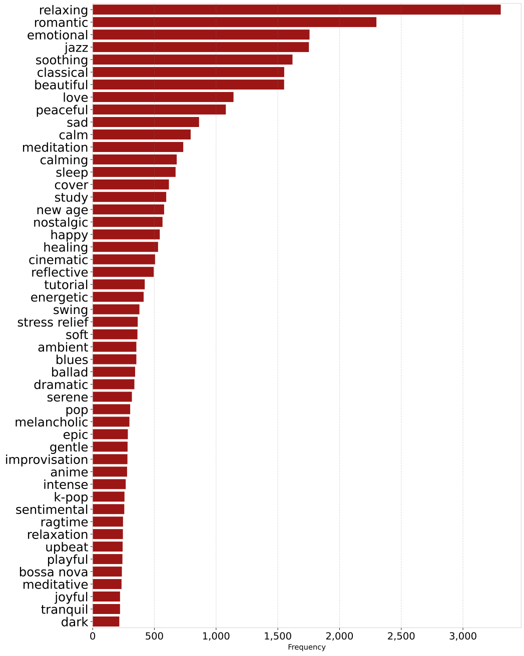

Figure 1: Top 50 words most frequently appearing in the text dataset of the PIAST-YT.

2.2 The PIAST-YT Dataset

The PIAST-YT dataset comprises approximately 9,673 tracks (1,006 hours) of audio collected from YouTube, accompanied by text information (title, tags, and descriptions of the video). We employed two collection methods: tag-based and channel-based. The tag-based method used our taxonomy to gather diverse styles of piano music from YouTube. However, the inherent variability in the availability of solo piano content on the platform led to some imbalance in the collected data. To ensure the inclusion of high-quality solo piano content, we also employed a channel-based method, collecting piano performance videos from 23 selected channels known for their piano content. Finally, The PIAST-YT dataset comprises three main components after pre-processing step: audio extracted from videos, text data (titles, tags, and descriptions), and MIDI data generated through transcription.

2.2.1 Pre-processing

Audio: To isolate pure piano solo performances, we filtered the data using musicnn (Pons and Serra, 2019), excluding tracks with non-piano sounds in their top 5 tags. Files exceeding 2 hours were removed, and those exceeding 30 minutes were segmented into 10-minute chunks for data consistency. This process reduced the original 1,789 hours of data, about 44%, to 1,006 hours.

Text: The collected text data from YouTube contained diverse and irrelevant information. To extract relevant music-descriptive features, we employed an LLM-based model, specifically ChatGPT 4-Turbo (Ouyang et al., 2022), chosen for its

[1] https://hayeonbang.github.io/PIAST_dataset/
[2] https://github.com/Hayeonbang/PIAST

high performance. This model generated a tag list for each video based on its corresponding text. Figure 1 illustrates the distribution of these generated tags. The total number of vocabulary is 3,160.

MIDI: The piano audio files were transcribed to performance MIDI using an automatic piano transcription model (Kong et al., 2021). The MIDI was then synchronised to beat estimates, and melody and chords were extracted using the Pop1k7 dataset pipeline (Hsiao et al., 2021). For the beat estimates, following (Holzapfel et al., 2012), we used the Mean Mutual Agreement (MMA) between a 'committee' of several state-of-the-art beat trackers, including All-in-One (Kim and Nam, 2023) and madmom (Böck et al., 2014), to filter out samples for which the beat tracking quality was poor. This transcription process was applied to the audio in both the PIAST-YT and the PIAST-AT datasets.

2.3 The PIAST-AT Dataset

Even after processing, the text data in the PIAST-YT exhibited several limitations. Although it was processed using an LLM-based model, it still showed a low correlation with the music content, and some audio files lacked corresponding text data. To address these issues, we created the PIAST-AT, a dataset consisting of piano-specific human-annotated text.

2.3.1 Annotation Process

We stratified 2,400 samples from audio data of PIAST-YT based on the queries used during the collection process, and extracted a 30-second segment from each sample for human annotation. The process involved 15 music experts (7 jazz and 8 classical musicians) with majors in composition. Each segment was assigned to three annotators using a web-based system. The annotators were divided into five groups tagged with 230 segments. Detailed descriptions and examples were provided to all annotators for each tag to ensure consistency. They were also instructed to exclude samples that did not strictly adhere to solo piano criteria or had subtle mood changes. After two rounds of annotation, we collected tags for 2,023 samples (approximately 17 hours of original audio), with 377 samples excluded through this process.

2.3.2 Dataset Analysis & Tag Consensus

Figure 2 shows the distribution of tags in the PIAST-AT dataset, categorized into Mood/Emotion, Genre, and Style. The Style category includes tags associated with performance difficulty and tempo-related mood. Due to the inherent imbalance of collected audio, there is also a disparity in the frequency of sub-genre tags.

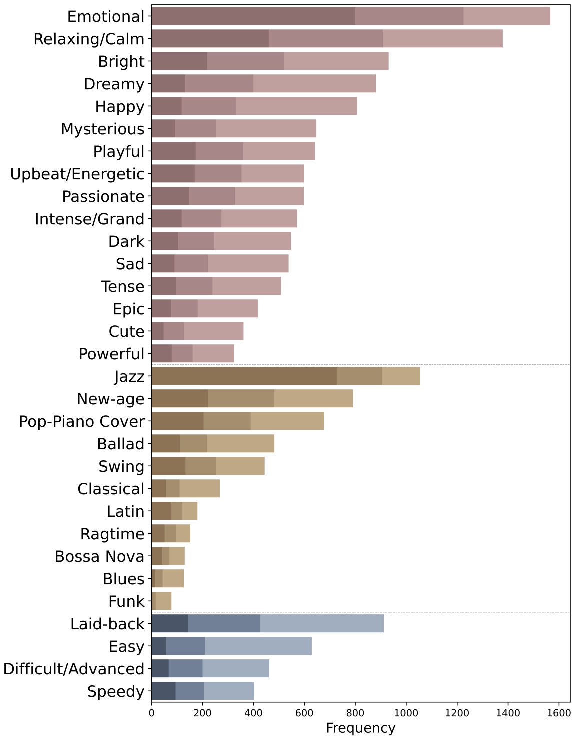

Figure 2: Tag distribution of the PIAST-AT dataset. Three distinct represent the degree of consensus. (Darkest: n=3, Medium: n=2, Lightest: n=1)

The dataset contains the consensus degree among the annotators. To leverage this information, we generated hierarchical captions based on the level of agreement as follows:

> *"This is definitely Jazz genre; (3 agreements)*
> *also Speedy style; also Playful mood; (2 agreements)*
> *potentially Latin genre; potentially Easy style; potentially Bright, Happy, Cute mood of piano music. (1 agreement)"*

The PIAST-AT dataset comprises audio, transcribed MIDI, and text annotations (tags and captions), offering a rich representation of musical characteristics and annotator consensus.

3 Piano Music Classification

In this section, we present the application of our proposed dataset for piano music annotation and retrieval tasks in both the audio and MIDI domains. We employed a two-stage framework: 1) pre-training and 2) transfer learning. For pre-training, we used the PIAST-YT dataset to train a general-purpose piano-specific model with large-scale audio, MIDI, and diverse text data. We leveraged text supervision through music-text joint embedding pre-training (Huang et al., 2022; Manco et al., 2022; Doh et al., 2024b). In the transfer learning stage, we utilized the PIAST-AT dataset to train a piano classification model as a downstream task.

3.1 Pre-training and Transfer Learning

To develop a piano-specific pre-trained model, we extracted embeddings from audio, MIDI, and text modality encoders. We applied contrastive loss to maximize similarity between corresponding pairs (audio-text or MIDI-text) while minimizing similarity with in-batch negative samples. Following previous studies (Huang et al., 2022; Manco et al., 2022; Doh et al., 2024b), each encoder consists of a modality-specific backbone, a linear projection layer, and an l_2 normalization layer. We used a modified ResNet-50 (Radford et al., 2021) for audio, RoBERTa (Liu et al., 2019) for text, and MidiBERT-Piano (Chou et al., 2021) with average pooling for MIDI.

For the classification model, we employed the probing protocol (Doh et al., 2023; Castellon et al., 2021). We used the pre-trained audio and MIDI encoders as frozen feature extractors and trained linear models and one-layer MLPs as shallow classifiers on top of them, with 512 hidden states and ReLU activation.

3.2 Implementation Details

We processed the input data for pre-training and transfer learning as follows: Audio inputs were 10-second signals sampled at 22050 Hz, converted to log-mel spectrograms with 128 mel bins using a 1024-point FFT with a Hann window and a 10 ms hop size. For MIDI, pre-processed MIDI files were converted to the CP (*compound word*) representation (Hsiao et al., 2021) and fed into a 12-layer BERT with a maximum sequence length of 512. All models were optimized using AdamW with a 5e-5 learning rate, and a dropout rate of 0.4 applied to the audio classification model. We used batch sizes of 128 for audio and 48 for MIDI data during pre-training. Pre-training models were trained for 150 epochs, while classification models ran for 700 epochs, with the best model selected based on validation loss. The PIAST-AT dataset was split into 80% for training, 10% for validation, and 10% for testing sets.

	Music → Tag		**Tag → Music**	
	ROC-AUC	**PR-AUC**	**ROC-AUC**	**PR-AUC**
Supervised				
Audio	81.06	70.71	73.22	50.97
MIDI	84.82	75.24	79.14	58.00
Pre-train and Transfer Learning				
Audio	84.52	74.73	79.01	58.70
MIDI	85.69	76.27	80.63	61.53

Table 1: Performance results for music-to-tag and tag-to-music tasks.

3.3 Evaluation & Results

We evaluated our classification models on two tasks: the *annotation task*, which involves finding appropriate tags for given music, and the *retrieval task*, which focuses on finding suitable music for provided tags. Following previous studies (Choi et al., 2019; Doh et al., 2024a), we employed the area under the ROC and PR curves averaged over instances as evaluation metrics for the annotation task. The retrieval task was assessed using the area under the ROC and PR curves averaged over labels. To demonstrate the effectiveness of the proposed large PIAST-YT dataset, we used a supervised model trained exclusively on the smaller PIAST-AT dataset as the baseline model.

Table 1 compares the annotation and retrieval performance across 1) audio and MIDI modalities, and 2) the supervised versus pre-train and transfer framework. The MIDI model consistently outperformed the audio model across both tasks. Pre-training with PIAST-YT improved the performance of both models on all metrics, demonstrating its effectiveness. This pre-training approach led to superior performance in both music-to-tag and tag-to-music tasks.

4 Conclusion

In this paper, we introduced PIAST, a piano dataset with audio, symbolic and text. Our experiments demonstrated the dataset's effectiveness for piano music annotation and retrieval tasks, showing improvements with pre-training. The PIAST dataset supports various applications, including improved music retrieval, text-based music generation, music analysis, and emotion/genre classification. To further enhance the dataset, our future work will address tag imbalances by adding more samples and incorporating additional processed data such as lead sheets and chord annotations.

Acknowledgments

This work was supported by the collaboration with NCSOFT, Korea.

References

Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. 2017. Developing a benchmark for emotional analysis of music. *PloS one*, 12(3):e0173392.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*.

Sebastian Böck, Florian Krebs, and Gerhard Widmer. 2014. A multi-model approach to beat tracking considering heterogeneous music styles. In *Proceedings of 15th International Conference on Music Information Retrieval (ISMIR)*.

Rodrigo Castellon, Chris Donahue, and Percy Liang. 2021. Codified audio language modeling learns useful representations for music information retrieval. In *Proceedings of the 22th International Society for Music Information Retrieval Conference (ISMIR)*.

Eunjin Choi, Yoonjin Chung, Seolhee Lee, JongIk Jeon, Taegyun Kwon, and Juhan Nam. 2022. YM2413-MDB: A multi-instrumental fm video game music dataset with emotion annotations. In *Proceedings of 23th International Conference on Music Information Retrieval (ISMIR)*.

Jeong Choi, Jongpil Lee, Jiyoung Park, and Juhan Nam. 2019. Zero-shot learning for audio-based music classification and tagging. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*.

Jongho Choi and Kyogu Lee. 2023. Pop2Piano: Pop audio-based piano cover generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yi-Hui Chou, I Chen, Chin-Jui Chang, Joann Ching, Yi-Hsuan Yang, et al. 2021. MidiBERT-Piano: Large-scale pre-training for symbolic music understanding. *arXiv preprint arXiv:2107.05223*.

SeungHeon Doh, Jongpil Lee, Dasaem Jeong, and Juhan Nam. 2024a. Musical word embedding for music tagging and retrieval. *arXiv preprint arXiv:2404.13569*.

SeungHeon Doh, Minhee Lee, Dasaem Jeong, and Juhan Nam. 2024b. Enriching music descriptions with a finetuned-llm and metadata for text-to-music retrieval. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 826–830. IEEE.

SeungHeon Doh, Minz Won, Keunwoo Choi, and Juhan Nam. 2023. Toward universal text-to-music retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Drew Edwards, Simon Dixon, and Emmanouil Benetos. 2023. Pijama: Piano jazz with automatic midi annotations. *Transactions of the International Society for Music Information Retrieval*.

Lucas N Ferreira and Jim Whitehead. 2019. Learning to generate music with sentiment. In *Proceedings of 20th International Conference on Music Information Retrieval (ISMIR)*.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2019. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *Proceedings of 7th International Conference on Learning Representations (ICLR)*.

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

Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. 2021. Compound Word Transformer: Learning to compose full-song music over dynamic directed hypergraphs. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, volume 35, pages 178–186.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. 2022. MuLan: A joint embedding of music audio and natural language. In *Proceedings of 23th International Conference on Music Information Retrieval (ISMIR)*.

Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. 2021. EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In *Proceedings of 22th International Conference on Music Information Retrieval (ISMIR)*.

Taejun Kim and Juhan Nam. 2023. All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE.

Qiuqiang Kong, Bochen Li, Jitong Chen, and Yuxuan Wang. 2022. GiantMIDI-Piano: A large-scale midi dataset for classical piano music. *Transactions of the International Society for Music Information Retrieval*, 5(1):87–98.

Qiuqiang Kong, Bochen Li, Xuchen Song, Yuan Wan, and Yuxuan Wang. 2021. High-resolution piano transcription with pedals by regressing onset and offset times. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3707–3717.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022. Contrastive audio-language learning for music. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jordi Pons and Xavier Serra. 2019. musicnn: Pretrained convolutional neural networks for music audio tagging. In *Late Breaking/Demo in the 20th International Society for Music Information Retrieval (ISMIR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763.

Amirreza Rouhi, Micol Spitale, Fabio Catania, Giulia Cosentino, Mirko Gelsomini, and Franca Garzotto. 2019. Emotify: emotional game for children with autism spectrum disorder based-on machine learning. In *Companion Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 31–32.

Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. 2007. Towards musical query-by-semantic-description using the cal500 data set. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 439–446.

Shih-Lun Wu and Yi-Hsuan Yang. 2023. Compose & Embellish: Well-structured piano performance generation via a two-stage approach. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.