IMPROVING AUTOMATIC DRUM TRANSCRIPTION USING LARGE-SCALE AUDIO-TO-MIDI ALIGNED DATA

I-Chieh Wei[1], *Chih-Wei Wu*[2], *Li Su*[1]

[1]Institute of Information Science, Academia Sinica, Taiwan
[2]Netflix, Inc., USA

ABSTRACT

One of the major challenges in Automatic Drum Transcription (ADT) research is the lack of large-scale labeled dataset featuring audio with polyphonic mixtures; this limitation around data availability greatly impedes the progress of data-driven approaches in the context of ADT. To tackle this issue, we propose a semi-automatic way of compiling a labeled dataset using the audio-to-MIDI alignment technique. The resulting dataset consists of 1565 polyphonic mixtures of music with audio-aligned MIDI ground truth. To validate the quality and generality of this dataset, an ADT model based on Convolutional Neural Network (CNN) is trained and evaluated on several publicly available datasets. The evaluation results suggest that our proposed model, which is trained solely on the compiled dataset, compares favorably with the state-of-the-art ADT systems. The result also implies the possibility of leveraging audio-to-MIDI alignment in creating datasets for a broader range of audio related tasks.

Index Terms— automatic drum transcription, audio-to-MIDI alignment, semi-supervised labeling

1. INTRODUCTION

Automatic Drum Transcription (ADT) [1] is a task that involves the isolation and identification of percussive events from audio signals. Similar to other sub-tasks in Automatic Music Transcription (AMT) [2], a general trend of adopting data-driven approaches such as Deep Neural Networks (DNNs) could be observed in ADT literature over the past years [3–8]. This increasing popularity of data-driven systems in ADT has led to the discussion concerning data insufficiency [1], and different labeled datasets such as MDB-Drums [9] and RBMA13 [10] have been introduced to the ADT community in order to address this issue directly. However, the high cost associated with the manual annotation process makes it very difficult to scale. As a result, these datasets are often limited in size and represent only a small portion of the real problem space. To make further progress without the intensive labor of creating labeled datasets, various approaches have been proposed. For instance, Wu and Lerch proposed to use unlabeled data for training a DNN model via teacher-student learning paradigm [11], and Choi and Cho [12] presented an ADT model that can be trained using unlabeled data via unsupervised learning. In addition to unlabeled data, the idea of using synthetic data generated from MIDI sequences has been explored by Vogl et al. [5] and Cartwright et al. [6]. Similarly, Callender et al. [13] proposed to build a dataset using electronic drum kits and synthetically extend the dataset with audio samples from multiple drum machines. Other techniques, such as data augmentation, have also been investigated and shown effective in the context of ADT [7]. Generally speaking, the above-mentioned approaches are able to provide marginal improvements when training resources are lacking, however, they still cannot fully replace the vital role of a labeled dataset.

Recently, an interesting idea of leveraging MIDI and audio data to create a large labeled dataset has been proposed for piano transcription [14]. In particular, the audio-to-MIDI alignment technique [15], along with other automated processes, has been applied to clean up an extensive collection of slightly misaligned MIDI and audio sequences. The resulting dataset was able to support the training of an advanced DNN architecture, leading towards a promising result.

Inspired by the above-mentioned studies, we explore the similar idea of extending an existing MIDI dataset for ADT use cases. The objective is to compile a sizable ADT dataset with real-world audio recordings and detailed annotations without inducing the intensive labors from the human annotators. Furthermore, we propose an ADT model that leverages the state-of-the-art beat tracker [16] in order to make beat-informed predictions. The contributions of this work include: (i) a large labeled dataset that contains audio recordings of 1,565 popular music with audio-aligned MIDI drum sequences, (ii) a beat-informed CNN model with attention mechanism for ADT that outperforms the state-of-the-art system, and (iii) an advanced semi-automatic procedure to create labeled datasets that could potentially be applied to other audio related tasks.

2. METHODOLOGY

The system overview is shown in Fig. 1. There are three major stages in our proposed method. In the data creation stage,

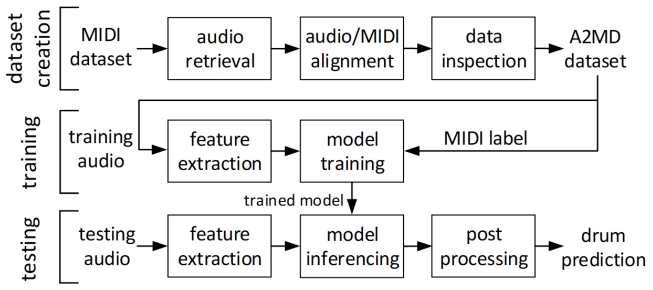

Fig. 1. Illustration of the proposed drum transcription system. The three fundamental stages are: (i) dataset creation, (ii) training, and (iii) testing. Further explanation of every single stage is presented in Section 2.

Name	**# Tracks**			**Total Dur. (hr)**
ENST	64			1.28
MDB-Drums	23			0.23
RBMA13	30			1.67
A2MD (Proposed)	**Subset**	**Alignment Level**	**# Tracks**	**Total Dur. (hr)**
	L	Low	574	34.5
	M	Medium	794	
	H	High	177	

Table 1. A comparison between drum datasets. Note that the size of A2MD is significantly larger than the other existing datasets in terms of the total duration.

a labeled dataset is created from a large existing MIDI dataset; this process includes the retrieval of the audio recordings that correspond to the MIDI data, the alignment between the audio and the MIDI sequences, and an inspection process for ensuring the data integrity. The resulting Audio-to-MIDI Drum (A2MD) dataset is subsequently used in the training stage. During training, the feature representations are extracted from the audio, and an ADT model is trained using the extracted features and the MIDI ground truth via supervised learning. In the testing stage, a similar pipeline is followed by extracting the features and deriving the activation functions of three drum instruments (i.e., kick drum, snare drum, hihat) using the trained model from the previous stage. Finally, the activation functions are post-processed, and the onsets of each drum event are returned as the final results. More details are elaborated in the following sections.

2.1. Dataset creation

In this work, we use the Lakh MIDI dataset [17] to build a large labeled dataset for ADT tasks; the Lakh dataset is chosen for its large and diverse collection of MIDI sequences. The original Lakh dataset contains the basic audio-to-MIDI alignment information (e.g., alignment score), however, it has the following drawbacks: (i) the audio files are not available and (ii) most of the drum sequences are not well-aligned. To address these issues, we propose the following steps to filter and clean up the Lakh dataset for ADT use cases:

- **Audio retrieval**: to retrieve the audio files that match the MIDI data, we first compile a list of songs with higher alignment scores as reported by Raffel [17]; this list allows us to ignore the unreliable data and focus on the songs with higher probabilities of being accurately aligned. Next, we identify the artist names and titles of these songs from the metadata; these attributes are used for querying online platforms such as YouTube, and the audio files from the top-10 search results are collected and stored in the MP3 format with a bit rate of 256 kbps. To select the best matching audio from these 10 candidates, the pair-wise similarity scores between the MIDI-synthesized audio and the downloaded audio are computed, and the file with the highest similarity score is chosen.
- **Audio-to-MIDI alignment**: in this step, we apply the audio-to-MIDI alignment, an algorithm based on Dynamic Time Warping (DTW) [17], to align the MIDI ground truth with the selected audio file. Additionally, We modify the original algorithm to dynamically search for the optimal penalty value P in order to quantify the alignment result. Generally speaking, a higher P is associated with a relatively coarse alignment result. The pair of audio and MIDI sequence is considered a match if a lower P is returned after the alignment process. Finally, for the matched pairs, we adjust the MIDI sequence based on the warping path to align the drum events with the audio signal.
- **Data inspection**: as a final step, we apply an automated inspection process to ensure the validity of the alignment results. Particularly, we compute another alignment process between the isolated drum signal and the MIDI drum sequence; the drum signal is extracted from the polyphonic mixture using an existing source separator [18]. Any pair with a $P > 0.7$ is discarded. This threshold is chosen based on our initial data observations. Lastly, a quick spot-check is performed manually on the polyphonic mixture to ensure that the MIDI and the audio are identical song.

The resulting dataset (referred to as A2MD) consists of 1565 audio/MIDI pairs as shown in Table 1. According to the distribution of penalty values, they could be further divided into A2MD-L ($0.4 \leq P < 0.7$), A2MD-M ($0.2 \leq P < 0.4$), and A2MD-H ($P < 0.2$), which correspond to low, medium, and high alignment level, respectively.

2.2. Feature extraction

In this processing step, the Constant-Q Transform (CQT) spectrogram is calculated from input audio with a hop size of 5.8 ms and a spectral resolution of 12 bins per octave.

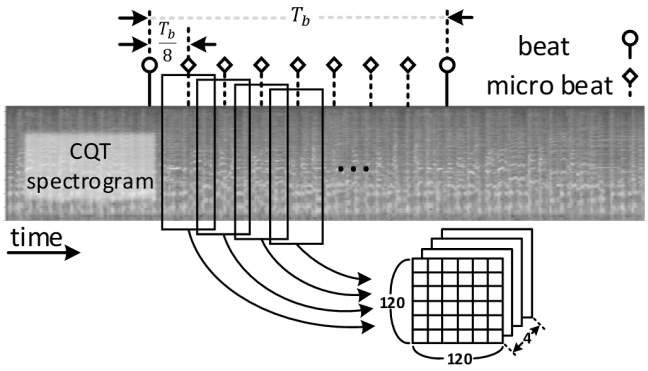

Fig. 2. Flowchart of the feature extraction process. The data processing procedure is presented in Section 2.2.

To prepare a musically meaningful input representation, we propose a beat-informed segmentation process as shown in Fig. 2. This process includes the following steps: first, all the beats and micro beats within the audio track are located using a state-of-the-art beat tracker [16]; each micro beat is a 32th note estimated by $T_b/8$, where T_b is the inter-beat interval. Next, a sliding window of 0.7 s is applied around each micro beat to divide the CQT spectrogram into overlapping segments; the size of this sliding window is chosen in order to cover the duration of most percussive events. Finally, for every four consecutive segments, a stacked representation with a dimensionality of $\{120 \times 120 \times 4\}$ is computed as the input to the ADT model.

2.3. CNN model architecture

We use a CNN-based model to predict the drum activation functions from the processed spectrogram that contains rich beat information. This beat-informed CNN model, referred to as B-CNN, is comprised of five convolutional layers, one attention layer, and three fully-connected layers. Each convolutional layer has 64 filters with the kernel size and stride equal to (3,3) and (1,1), respectively. The number of neurons in the fully-connected layer is set to 1,024. For the attention layer, we used the same design as in [19] with minor adjustments to the input/output dimensionality. The B-CNN model is trained to minimize the Mean Absolute Error (MAE) between model output $\mathcal{V}_{pred}$ and groundtruth label $\mathcal{V}_{label}$ from MIDI drum sequence. The loss function can be expressed as follows:

$$\mathcal{L}_{drum} = \text{MAE}\left\{\mathcal{V}_{pred}, \mathcal{V}_{label}\right\}, \tag{1}$$

where $\{\mathcal{V}_{pred}, \mathcal{V}_{label}\} \in R^{m \times n}$. m is the number of drum instruments and n is the duration in terms of micro beat. In this work, we set $m = 3$ and $n = 4$. The B-CNN model is implemented using Tensorflow[1]. All the weights are randomly initialized and optimized using Adam [20] optimizer with a batch size of 128 and a fixed learning rate of 2e-5. The total number of parameters of B-CNN is around 9.4M. While training on a system with single 1080 Ti GPU, it takes around 42 minutes to finish one epoch (loop over the whole A2MD data). The training process is manually stopped after 15 epochs, and the required time is around 11 hours.

[1]https://www.tensorflow.org/, last accessed: 2020.10.19

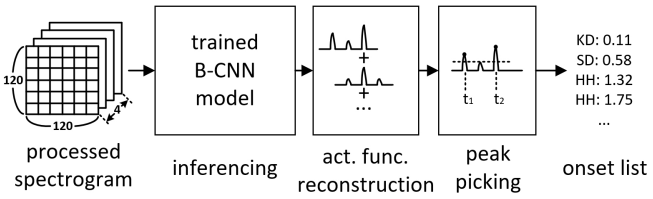

Fig. 3. Flowchart of the post-processing steps in the testing stage. The two applied steps are: (i) activation function reconstruction, and (ii) peak-picking. The segmented activation function generated by B-CNN is reconstructed first. Then, a peak-picking function is applied to obtain onset peaks.

2.4. Post-processing

As shown in Fig. 3, there are two steps to identify drum onset locations from the activation functions generated by the B-CNN model. First, the activation function of each drum instrument is reconstructed based on the original time indices. Then, a peak-picking function, which is similar to the one used in the previous work [5, 6], is applied to locate the onset times for each drum instrument. The final output of the post-processing step is a list containing drum onset predictions.

3. EXPERIMENTAL SETUP

Two sets of experiments are conducted to investigate the effectiveness of our proposed B-CNN model and A2MD dataset. In the first set of experiment, we evaluate the performance of B-CNN model by comparing with the state-of-the-art systems (i.e., the CNN and CRNN as described in [5]). To ensure the compatibility of the experiment results, we follow the procedure in [5] and apply 3-fold cross validation for each of the three public datasets. These datasets include ENST [21], MDB-Drums [9], and RBMA13 [10] (see Table. 1).

The second set of the experiment aims to evaluate the usability of the proposed A2MD dataset. To this end, the identical B-CNN model architecture is trained using three subsets of data with different alignment quality in A2MD. Specifically, the three models are trained using the high-quality subset (A2MD-H), high- and medium- quality subset (A2MD-H plus A2MD-M), and the entire dataset (A2MD) respectively. We then use three aforementioned public datasets (ENST, MDB-Drums, and RBMA13) for testing. It should

Test dataset	Systems					
	Trained with ENST/MDB-Drums/RBMA13			Trained with A2MD subsets		
	CNN [5] (3-fold)	CRNN [5] (3-fold)	B-CNN (3-fold)	B-CNN (H)	B-CNN (H+M)	B-CNN (H+M+L)
ENST	0.72	0.71	**0.86**	0.69	0.72	**0.74**
MDB-Drums	0.67	0.67	**0.70**	0.64	0.70	**0.71**
RBMA13	**0.65**	0.63	0.62	0.53	0.57	**0.58**

Table 2. Overall F-score of systems evaluated on different datasets. The models in the left half table are trained using existing datasets. The models in the right half table are trained using the combanations of A2MD subset. For the CNN and CRNN model, we directly report the numbers in previous paper. All the B-CNN models are implemented with the same DNN architecture.

be noted that we intentionally apply different datasets for training (the proposed A2MD) and testing (the three public datasets), which is the most generalizable use case.

The evaluation metrics are the conventional precision (P), recall (R), and F1-score (F). The tolerance window for onset matching is 50 ms. For each model, we compute the P, R, and F for three drum instruments (i.e., kick drum, snare drum, hihat). However, only the averaged F-scores across three instruments are reported here due to the space limitation. Please refer to our online repository for detailed reports[2].

4. RESULTS AND DISCUSSION

The results of the first set of experiment are shown in the left three columns of Table 2. It can be observed that our proposed B-CNN model outperforms other existing models (i.e., CNN and CRNN) both on ENST and MDB-Drums datasets. Particularly, an obvious improvement can be seen on the ENST dataset, which may be owing to the lack of data diversity in ENST. ENST possesses the highest intra-dataset similarity among all datasets, since it only includes music tracks played by three drummers. Our B-CNN built based on beat information has proven to be efficient in capturing the drum events in ENST, and even yielded the most favorable results against the state-of-the-art systems on other two datasets with higher diversity.

The results of the second set of experiment are shown on the right side of the Table 2. The key findings are: first, all of our B-CNN models can achieve comparable performances with the state-of-the-art systems for all subsets of training data. This result suggests the usefulness and generalizability of our proposed A2MD dataset for training ADT models. Second, the model performance improves gradually as more data are included for training. However, the improvement becomes marginal when data with lower alignment quality are added. This result suggests that further improvements could potentially be achieved by increasing the alignment quality of A2MD-M and A2MD-L. Finally, the B-CNN trained with the entire A2MD performs better on ENST and MDB-Drums datasets, compared to the 3-fold cross validation for CNN and CRNN. This improvement is worth noting, especially considering that the training data and testing data of B-CNN are from diverse datasets. This result highlights the advantage of having a sizable dataset with high content diversity.

It should be noted that the distinctive music style and instrumentation of RBMA13 may lead to the mediocre performance of B-CNN models when tested on the RBMA13 dataset. Specifically, the sound tracks in RBMA13 are generated by electronic drum kits or samplers, which are very different from the conventional acoustic drum sound in other datasets, including A2MD. This difference may pose additional challenges for the beat-tracker or the model, which causes the sub-optimal performance of B-CNN.

5. CONCLUSION

In this work, we presented a large labeled dataset and a beat-informed CNN model for ADT tasks. Based on our experiments, the proposed A2MD dataset is able to support the training of the B-CNN model and lead to promising performances. Additionally, the results suggest the applicability of our semi-automatic process for creating a large labeled dataset using our adaptation of the audio-to-MIDI alignment technique. It is worth noting that our proposed B-CNN requires the beat information from a beat-tracker for feature extraction. To quantify the potential impact from inaccurate beat-tracking results, more investigations would be required. In addition, further enhancement of the alignment procedure can be made to improve the quality of the generated dataset. Finally, we would like to increase the diversity of A2MD dataset and include more music genres. We believe this semi-automatic approach of creating large labeled dataset has the potential of enabling the creation of more datasets for other audio related tasks.

6. REFERENCES

[1] Chih Wei Wu, Christian Dittmar, Carl Southall, Richard Vogl, Gerhard Widmer, Jason Hockman, Meinard Mueller, and Alexander Lerch, "A review of automatic

[2] https://github.com/Sma1033/adt_with_a2md, last accessed: 2020.10.19

drum transcription," *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2018.

[2] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, 2013.

[3] Richard Vogl, Matthias Dorfer, Gerhard Widmer, and Peter Knees, "Drum Transcription Via Joint Beat and Drum Modeling Using Convolutional Recurrent Neural Networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[4] Carl Southall, Ryan Stables, and Jason Hockman, "Automatic drum transcription for polyphonic recordings using soft attention mechanisms and convolutional neural networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[5] Richard Vogl, Gerhard Widmer, and Peter Knees, "Towards multi-instrument drum transcription," in *Proceedings of International Conference on Digital Audio Effects (DAFx)*, 2018.

[6] Mark Cartwright and Juan Pablo Bello, "Increasing drum transcription vocabulary using data synthesis," in *Proceedings of International Conference on Digital Audio Effects (DAFx)*, 2018.

[7] Celine Jacques and Axel Röbel, "Data augmentation for drum transcription with convolutional neural networks," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2019.

[8] Shun Ueda, Kentaro Shibata, Yusuke Wada, Ryo Nishikimi, Eita Nakamura, and Kazuyoshi Yoshii, "Bayesian drum transcription based on nonnegative matrix factor decomposition with a deep score prior," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2019.

[9] Carl Southall, Chih-Wei Wu, Alexander Lerch, and Jason Hockman, "Mdb drums: An annotated subset of medleydb for automatic drum transcription," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR) Late-breaking and Demo Papers*, 2017.

[10] Richard Vogl, Matthias Dorfer, Gerhard Widmer, and Peter Knees, "Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks.," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[11] Chih-Wei Wu and Alexander Lerch, "Automatic drum transcription using the student-teacher learning paradigm with unlabeled music data," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[12] Keunwoo Choi and Kyunghyun Cho, "Deep unsupervised drum transcription," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[13] Lee Callender, Curtis Hawthorne, and Jesse Engel, "Improving perceptual quality of drum transcription with the expanded groove midi dataset," *arXiv:2004.00188*, 2020.

[14] 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 (ICLR)*, 2019.

[15] Colin Raffel and D. Ellis, "Large-scale content-based matching of midi and audio files," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[16] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer, "madmom: a new Python Audio and Music Signal Processing Library," in *Proceedings of the 24th ACM International Conference on Multimedia (ACMMM)*, 2016.

[17] Colin Raffel, *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*, Ph.D. thesis, Columbia University, New York, NY, 2016.

[18] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach, "Music source separation in the waveform domain," *arXiv:1911.13254*, 2019.

[19] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning (ICML)*, 2019.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2014.

[21] Olivier Gillet and Gaël Richard, "Enst-drums: an extensive audio-visual database for drum signals processing.," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2006.