ENHANCED AUTOMATIC DRUM TRANSCRIPTION VIA DRUM STEM SOURCE SEPARATION

Xavier Riley
Centre for Digital Music
j.x.riley@qmul.ac.uk

Simon Dixon
Centre for Digital Music
Queen Mary University of London

ABSTRACT

Automatic Drum Transcription (ADT) remains a challenging task in MIR but recent advances allow accurate transcription of drum kits with up 5 classes - kick, snare, hi-hats, toms and cymbals - via the ADTOF package. In addition, several drum kit *stem* separation models in the open source community support separation for more than 6 stem classes, including distinct crash and ride cymbals. In this work we explore the benefits of combining these tools to improve the realism of drum transcriptions. We describe a simple post-processing step which expands the transcription output from five to seven classes and furthermore, we are able to estimate MIDI velocity values based on the separated stems. Our solution achieves strong performance when assessed against a baseline of 8-class drum transcription and produces realistic MIDI transcriptions suitable for MIR or music production tasks.

1. INTRODUCTION

Automatic Drum Transcription (ADT), a sub-task of Automatic Music Transcription (AMT), offers huge potential for extracting useful data from music signals. The history of ADT methods is well summarised in the literature [1,2]. Both cited surveys show that the history of ADT mirrors that of AMT in terms of which techniques were adopted; for example the progression from signal processing techniques to NMF based approaches to the more recent deep learning methods. However, while AMT accuracy for pitched instruments has increased steadily over time [3–5], the accuracy for ADT methods remains far below that of, say, piano. As identified by Wu et al. [2], they are affected by data issues such as small size, lack of complexity or lack of diversity (homogeneity).

Callender et al. [6] showed that an Onsets and Frames [3] style ADT model which included velocity data demonstrated strong performance in user preference studies, due to the inclusion of velocity information. However, the released model appears to struggle with generalisation as MP3 encoding can negatively affect real-world transcription results [1] .

Vogl et al. [7] introduced the use of a Convolutional Recurrent Neural Network (CRNN) which transcribed 3 drum classes (BD, SN, HH) with SOTA results on the ENST dataset. This was later expanded to 8 drum classes [8], albeit with lower accuracy. The current state of the art for ADT is ADTOF [9], which uses a similar CRNN architecture but with scaled up training data. However, the output of the ADTOF model is limited to 5 classes due to the source material used as training data.

Source separation has seen an explosion of interest in recent years with the advent of deep learning models for this task. While separating drum kits from a mixture is well studied [10], a relatively new task is performing audio separation of individual components from a drum kit performance - for example decomposing a drum kit into kick, snare, toms, hi-hat and cymbal stems. LarsNet [11] is one such solution, however the training data is derived entirely from drum samples produced by Logic Pro X. This lack of diversity could harm separation quality when extending to real-world examples.

Our work explores a recent open source contribution, which we refer to as "Jarredou" [2] , that has not yet been described in the scientific literature. In correspondence with the authors, we understand that this is trained on a private dataset of MIDI and rendered audio from drum-sample libraries with 21.8 hours of audio in the training set and 0.27 hours of audio reserved as a validation set. There is some repetition in the MIDI annotations as they are re-rendered using a variety of sample libraries from different providers. A full breakdown of its separation performance is planned for future work, however empirically we find that it performs well on a variety of recordings.

2. METHOD

Our method operates on solo drum kit audio (solo drums). Where the desired source is already part of a mix, we first isolate the drum part using Demucs v4 [10]. The input audio is initially normalized to a constant level using the ReplayGain algorithm [3] to ensure a relatively consistent dynamic level for later processing. We proceed to transcribe the solo drums via ADTOF to extract note locations for 5 drum classes (kick, snare, hi-hat, toms and cymbals).

[1] github.com/magenta/magenta/issues/1876

[2] github.com/jarredou/models/releases/tag/aufr33-jarredou_MDX23C_DrumSep_model_v0.1

[3] implemented in Essentia [12]

Late-Breaking / Demo Session Extended Abstract, ISMIR 2024 Conference

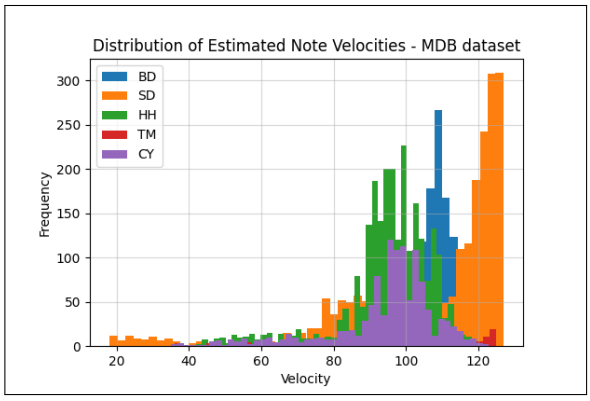

Figure 1. Estimated Velocity Distributions for the MDB dataset

We also separate the solo drums using the Jarredou model to extract 6 stems, with the cymbals class is expanded to crash and ride stems.

To recover velocity information we marry the transcription data to the drum stem data as follows: first a loudness curve is computed for each stem. We use an equal loudness filter from Essentia to preserve perceptual balance, then calculate the RMS with a 1024 sample window (at a 44.1kHz sample rate) and a 10ms hop size. The RMS values are then converted to a decibel (log) scale. All six stems are then normalized to the peak dB across the group – this allows the performance velocity to be self-consistent within a single performance, but if a reference level is known it can be substituted instead.

For each note in the transcribed MIDI, we take a 50ms window around the predicted onset time applied to the loudness curve for the corresponding stem. We then extract a MIDI velocity estimate from 0-127 by taking the maximum value of the window, scaled to the normalized dynamic range across all stems in a performance.

2.1 Recovering additional instrument classes

We further enhance the transcription by increasing the number of predicted classes from 5 to 7 as follows. For the cymbal onset predictions, we compare the loudness curves for the crash and ride stems and choose the maximum. However, the nature of crash cymbals means they often have a long, slow decay in amplitude which can cause incorrect classifications. To address this we introduce a heuristic by identifying significant crash cymbal peaks over the entire performance. For each crash peak, a refraction period is added which lasts until 1 second before the next peak. During this period a crash cymbal cannot be re-triggered and all cymbal hits during the period are assigned to the ride instead.

For the hi-hat stem, we observe that the loudness curve for open hi-hats decays more slowly. For each hi-hat note we take a window of the loudness curve up to the next hi-hat onset or 150ms, whichever is smaller. If the minimum loudness over the window is greater than 75% of the maximum we assign the note to the open hi-hat class, otherwise it is assigned to the closed hi-hat class.

Method	Dataset	Recall	Precision	F-measure
Ours	MDB	0.89	0.89	0.89 *(0.87)*
Ours	RBMA†	0.62	0.66	0.63 *(0.65)*
Ours	ENST	0.81	0.91	0.85 *(0.84)*

Table 1. Results for 5-class transcription accuracy across 3 datasets. Onset-only with 50ms tolerance via `mir_eval`. Obelisk (†) indicates drums were isolated using [10]. Original ADTOF results are shown in parentheses

Method	Dataset	Recall	Precision	F-measure
Ours	MDB	0.84	0.84	0.84 *(0.72)*
Ours	RBMA†	0.55	0.60	0.56 *(0.58)*
Ours	ENST	0.72	0.81	0.76 *(0.65)*

Table 2. Results for 8-class transcription accuracy. See Table 1 for details. Results in parentheses show baseline results for 8-class transcription from Vogl et al. [8].

3. RESULTS

The most commonly used datasets for ADT evaluation to date are ENST [13], MDB [14] and RBMA [7] which are all publicly available. Table 1 shows results for transcribing 5 drum classes (BD, SN, HH, CY, TOMS) following the framework of Zehren et al. [9]. The underlying transcription accuracy is essentially the same as that of ADTOF, with the only difference being our treatment of zero velocity notes which are omitted from the evaluation.

We also include results for the 8-class transcription task in Table 2. Our model does not attempt to predict the relatively rare Cowbell class (MIDI 56) so a direct comparison is not possible, however we include baseline results [8] for reference. These show strong increases in performance over the baseline for the MDB and ENST datasets (12% and 10%), however RBMA is 2% lower. We believe this is due to the electronic drum sounds used in this dataset which are outside of the domain of acoustic drums.

The predicted velocity distributions for the MDB dataset are shown in Figure 1. The other datasets are omitted but show similar distributions. This illustrates that our method produces a range of normally distributed velocities for each class. The concentration of snare at high volumes is likely a result of the equal loudness curves favouring mid-range frequencies. Our method allows for these components to be scaled individually to improve balance, if necessary.

4. CONCLUSIONS

In this work we demonstrate a method of combining an ADT model (ADTOF) with a drum stem source separation model. This combination allows us to estimate velocities and perform additional levels of classification while retaining a high degree of transcription accuracy. We intend to use this in future for dataset production workflows to enhance ADT further.

5. ACKNOWLEDGMENTS

XR is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported by UK Research and Innovation [grant number EP/S022694/1].

6. REFERENCES

[1] D. FitzGerald and J. Paulus, "Unpitched Percussion Transcription," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. Springer US, 2006, pp. 131–162.

[2] C. Wu, C. Dittmar, C. Southall, R. Vogl, G. Widmer, J. Hockman, M. Müller, and A. Lerch, "A review of automatic drum transcription," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 9, pp. 1457–1483, 2018.

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

[4] 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.

[5] Y. Yan and Z. Duan, "Scoring intervals using non-hierarchical transformer for automatic piano transcription," *CoRR*, vol. abs/2404.09466, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2404.09466

[6] L. F. Callender, C. G.-M. Hawthorne, and J. Engel, "Improving perceptual quality of drum transcription with the expanded groove midi dataset," *ArXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2004.00188

[7] R. Vogl, M. Dorfer, and P. Knees, "Drum transcription from polyphonic music with recurrent neural networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. IEEE, 2017, pp. 201–205.

[8] R. Vogl, G. Widmer, and P. Knees, "Towards multi-instrument drum transcription," *CoRR*, vol. abs/1806.06676, 2018. [Online]. Available: http://arxiv.org/abs/1806.06676

[9] M. Zehren, M. Alunno, and P. Bientinesi, "High-quality and reproducible automatic drum transcription from crowdsourced data," *Signals*, vol. 4, no. 4, pp. 768–787, 2023.

[10] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *ICASSP 23*, 2023.

[11] A. I. Mezza, R. Giampiccolo, A. Bernardini, and A. Sarti, "Toward deep drum source separation," *Pattern Recognition Letters*, vol. 183, pp. 86–91, 2024.

[12] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "Essentia: an open-source library for sound and music analysis," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 855–858.

[13] O. Gillet and G. Richard, "ENST-Drums: an extensive audio-visual database for drum signals processing," in *ISMIR 2006, 7th International Conference on Music Information Retrieval, Victoria, Canada, 8-12 October 2006, Proceedings*, 2006, pp. 156–159.

[14] C. Southall, C.-W. Wu, A. Lerch, and J. Hockman, "MDB Drums: An annotated subset of MedleyDB for automatic drum transcription," in *Late Breaking Demo at the 18th International Society for Music Information Retrieval Conference*, 2017.