Improving Perceptual Quality of Drum Transcription with the Expanded Groove MIDI Dataset

Lee Callender,[*] **Curtis Hawthorne,**[*] **Jesse Engel**

Google

leefcallender@gmail.com, fjord@google.com, jesseengel@google.com

Abstract

We introduce the Expanded Groove MIDI dataset (E-GMD), an automatic drum transcription (ADT) dataset that contains 444 hours of audio from 43 drum kits, making it an order of magnitude larger than similar datasets, and the first with human-performed velocity annotations. We use E-GMD to optimize classifiers for use in downstream generation by predicting expressive dynamics (velocity) and show with listening tests that they produce outputs with improved perceptual quality, despite similar results on classification metrics. Via the listening tests, we argue that standard classifier metrics, such as accuracy and F-measure score, are insufficient proxies of performance in downstream tasks because they do not fully align with the perceptual quality of generated outputs.

1 Introduction

Discriminative models predict the conditional distribution $p(y|x)$ over labels y that correspond to an input x. In the space of automatic drum transcription (ADT), discriminative models are used to predict when and what drum hits are used in a drum performance conditional on audio input of a performance.

While classifier metrics such as accuracy, precision, recall, and F-measure scores are often used to evaluate discriminative models, decision theory highlights that the true quantity of interest is the expected utility (or cost) of the inferred labels in a downstream task (Von Neumann, Morgenstern, and Kuhn 2007).

Recent work on piano transcription has demonstrated the value of considering downstream generation, showing that separately classifying note onsets from note persistence led to dramatic improvements in the perceptual quality of generation due to a reduction in false positive onsets (Hawthorne et al. 2018). For the application of drum transcription, we develop a new dataset and transcription model capable of transcribing drum hit velocity (loudness) and examine how that capability contributes to the perceived quality of the transcriptions.

Our key contributions include:

- The Expanded Groove MIDI dataset (E-GMD), the first dataset to capture both expressive timing and velocity of

human performances and a dataset size that is an order of magnitude larger than similar datasets.
- Training expressive ADT models on E-GMD to predict timings, drum hit, and velocity by incorporating a separate velocity-prediction head.
- Demonstrating that predicting expressive dynamics (velocity) in addition to timing generates outputs with improved perceptual quality, as determined by listening tests, despite achieving similar results on classification metrics.
- Developing a new *Shuffled mixup* strategy for data augmentation and regularization that effectively limits overfitting.

Audio samples of the dataset and examples used in the listening test are provided in the online supplement at https://goo.gl/magenta/e-gmd-examples, and the full dataset is available at https://g.co/magenta/e-gmd under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

2 Related Work

The recent work of Wu et al. (2018) provides a comprehensive overview of ADT and includes evaluation of current state of the art methods. While there has been a large collection of studies published over ADT in recent years (Vogl, Widmer, and Knees 2018; Choi and Cho 2019; Cartwright 2018; Wu and Lerch 2018; Southall, Stables, and Hockman 2018a,b; Ueda et al. 2019), most ADT research has maintained a focus on classifier metrics to assess quality.

Of the approaches that have explored deep learning (Vogl, Widmer, and Knees 2018; Choi and Cho 2019; Cartwright 2018; Southall, Stables, and Hockman 2018a), research is still fairly new given the large data required to effectively produce a model. As annotating drums is still a fairly manual task, most datasets for ADT are relatively small in size and resource intensive to create. This has lead to new research into solving that problem, including unsupervised approaches (Choi and Cho 2019; Wu and Lerch 2018) and the creation of synthetic datasets (Choi and Cho 2019; Vogl, Widmer, and Knees 2018; Cartwright 2018; Miron, Davies, and Gouyon 2013).

Given the difficulty of ADT and the limited datasets available, the overwhelming majority of ADT research has focused on ADT with the classification of 3 primary drum hits:

---

[*]Equal contribution

Kick Drum, Snare Drum, Hi-hat (KD, SN, HH) (Dittmar and Gärtner 2014a; Lindsay-Smith, McDonald, and Sandler 2012; Wu and Lerch 2015; Vogl, Dorfer, and Knees 2016, 2017; Stables, Hockman, and Southall 2016; Southall, Stables, and Hockman 2017). A handful of datasets contain annotations beyond the 3 standard hits, however the set of drum hits is not standardized, with each dataset containing a varied collection of drum hits (Vogl, Widmer, and Knees 2018; Cartwright 2018; Dittmar and Uhle 2004).

Velocity has sometimes been considered during ADT tasks. For example, in DrummerNet (Choi and Cho 2019), velocity is used as a probability of hit for peak-picking. However, velocity is not predicted as part of overall model output. To the best of our knowledge, our work is the first model that directly predicts velocity values and evaluates the perceptual quality of resynthesized outputs.

Dataset	Minutes	Kits	Human	Vel
E-GMD	26,670	43	$\checkmark$	$\checkmark$
TMIDT	15,540	57	$\times$	$\times$
IDMT	130	6	$\times$	$\times$
ENST	61	3	$\checkmark$	$\times$
MDB Drums	21	≈ 23	$\checkmark$	$\times$
RBMA13	103	≈ 30	$\checkmark$	$\times$

Table 1: Comparison of public datasets for ADT, including whether they contain exclusively human performances and velocity annotations. The exact number of kits in MDB Drums and RBMA13 is unclear, but is unlikely to exceed the total number of tracks, which is 23 and 30 respectively. All datasets contain isolated drum tracks, with the exception of RBMA13.

3 Datasets

Only a handful of public datasets are available for ADT, and many have limited size and diversity. An even smaller subset of datasets contain human performances, and no public datasets contain human performances with velocity annotations (Cartwright 2018; Wu et al. 2018; Vogl, Widmer, and Knees 2018). Reasons for these limitations include the tedious nature of generating labels for real drum performances and restrictions around licensing and intellectual property.

The difficulty of annotating real drum performances has inspired some recent studies to generate their own synthetic datasets. These datasets are commonly generated by taking a collection of MIDI (Music Instrument Digital Interface, the industry standard format for symbolic music data) drum performances and synthesizing audio via drum samples (Miron, Davies, and Gouyon 2013; Vogl, Widmer, and Knees 2018; Cartwright 2018). Only one of these datasets is public (Vogl, Widmer, and Knees 2018), and it does not contain velocity annotations.

Table 1 compares several public datasets, including E-GMD. Of these datasets, we decided to use IDMT-SMT (Dittmar and Gärtner 2014b) and ENST (Gillet and Richard 2006) in our evaluations because of their commonality in prior studies. We opted not to use MDB

Drums (Southall et al. 2017) because of its small size and did not use the dataset from Vogl et al. (2018), which we refer to as TMIDT, because the licensing of its source material was ambiguous. We also did not use RBMA13 (Vogl et al. 2017) because the tracks included music in addition to drums, and we focused on transcribing only solo drumming.

E-GMD has many different annotated hits. For evaluation and listening tests, we group the annotated hits down to a 7 and 3 hit classification task, as shown in Table 2.

E-GMD Hits	7 hit	3 hit
Kick drum	KD	KD
Snare drum	SD	
Snare rim		
Cross-stick		
Clap		SD
Tom 1	TT	
Tom 1 Rim		
Tom 2		
Tom 2 Rim		
Tom 3		
Tom 3 Rim		
Open Hi-Hat	HH	
Open Hi-Hat Bow		
Closed Hi-Hat Bow		
Closed Hi-Hat Bow		
Pedal Hi-Hat		
Tambourine		HH
Crash 1 Bow	CY	
Crash 1 Edge		
Crash 2 Bow		
Crash 2 Edge		
Ride Bow	RD	
Ride Edge		
Ride Bell	BE	
Cow Bell		

Table 2: The drum hit hierarchy for E-GMD. The 3 and 7 hit groupings are used in our model for evaluation and the listening test.

IDMT-SMT

IDMT-SMT contains only the 3 standard drum hits (KD, SN, HH), and contains 4 different drum kits. The dataset uses relatively simple drum patterns and contains audio and ground truth hit annotations. One drum kit is an acoustic kit that was recorded with varying velocities, however the ground truth annotations do not consider velocity and only consider drum hit type and timing. The other 3 drum kits use synthesized drums. The dataset contains audio for both individual hits and the mix of 3 hits. We use the full audio mix recordings for evaluation, and use the entire dataset because it is limited in length.

ENST

The ENST dataset was recorded with three different acoustic drum kits, performed by three professional drummers. Each

performer used either sticks, rods, brushes, or mallets for each sequence, to produce a variety of timbres.

The dataset contains audio of single instrument strokes, short phrases, and drum tracks with and without additional accompaniment. The annotations contain labels for 20 different drum hits. While the performances for ENST are recorded, there again is no velocity annotation.

For our experiments, the tracks of isolated drum performances were used (the tracks labeled "minus-one"), which is consistent with the other ADT studies we compare against. These isolated drum performances make up 64 tracks of 61s average duration and a total duration of 1 hour. We use all 64 tracks in evaluation. The rest of the dataset (single strokes, patterns) is ignored.

Expanded Groove MIDI Dataset

We introduce an expansion of the Groove MIDI Dataset (GMD), which we call the Expanded Groove MIDI Dataset (E-GMD). GMD is a dataset of human drum performances recorded in MIDI format on a Roland TD-11[1] electronic drum kit, and was originally created for generative drum sequencing (Gillick et al. 2019). MIDI information includes events like notes, that associate instrument, a time and a velocity together as an event.

GMD contains 13.6 hours, 1,150 MIDI files, and 22 different drum instruments. The dataset additionally includes synthesized audio outputs of the TD-11 aligned within 2ms of the corresponding MIDI files. The data includes performances by a total of 10 drummers, 5 professionals and 5 amateurs, with more than 80 percent coming from the professionals. The professionals were able to improvise in a wide range of styles, resulting in a diverse range of performances.

To make the dataset applicable to ADT, we expanded it by recording 43 drumkits on a Roland TD-17[2], ranging from electronic (e.g., 808, 909) to acoustic sounds. The additional drumkits were recorded at 44.1kHz and 24 bits and aligned within 2ms of the original MIDI files. Using the Roland TD-17, a close analog to the Roland TD-11 (no longer manufactured) used in the original Groove dataset, enables accurate reproduction of nuances in the initial performances.

We implemented a semi-manual process to systematically record new audio from the TD-17. The audio was recorded in real-time on a Digital Audio Workstation (DAW) and took about 16 hours to complete per kit. Given the semi-manual nature of the pipeline, there were some errors in the recording process that resulted in unusable tracks. The final numbers for E-GMD are shown in Table 3.

We maintained the same train, test and validation splits across sequences that GMD had. As each kit was recorded for every sequence, we see all 43 kits in the train, test and validation splits. The count of hits across all splits is shown in Table 4.

The online supplement includes examples of different sequences and kits at https://goo.gl/magenta/e-gmd-examples.

---

[1]https://www.roland.com/us/products/td-11/

[2]The TD-17 is an award-winning electronic drum kit that "faithfully reproduces the character and tone of acoustic drums." https://www.roland.com/us/products/td-17_series/

Split	Unique Seq	Total Seq	Dur
Train	819	35,217	341.4h
Test	123	5,289	50.9h
Validation	117	5,031	52.2h
Total	1,059	45,537	444.5h

Table 3: E-GMD unique sequences, total sequences, and duration in hours by split.

Hit	Train	Test	Validation
KD	2,181k	319k	343k
SD	3,477k	468k	533k
HH	3,045k	553k	518k
TT	805k	98k	171k
RD	1,260k	105k	84k
BE	191k	9k	21k
CY	122k	10k	27k

Table 4: E-GMD hit counts across splits in thousands. We show the counts for the seven hit grouping of E-GMD for brevity. See Table 2 for hit definitions and grouping description.

The dataset is available at https://g.co/magenta/e-gmd under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The model described in this paper was trained with the v1.0.0 release of the dataset.

4 Model

We base our model on Onsets and Frames (Hawthorne et al. 2018) and adapt its note and velocity prediction capabilities to drum hit and velocity predictions. We call our new model OaF-Drums.

We use only the onset and velocity stacks of the network, as illustrated in Figure 1, because drum hits do not sustain like piano notes and so we do not require the frame or offset predictions. Complete network details are given in the Supplement.
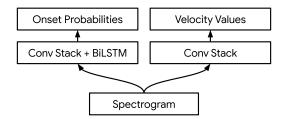

Figure 1: OaF-Drums Model Architecture

For log mel-spectrogram creation, we increased the audio sample rate from 16 KHz to 44.1 KHz, the number of bins from 229 to 250, and shortened the hop length from 512 to 441 samples, resulting in frames with a 10ms width (Vogl, Widmer, and Knees 2018). We found the higher sample rate improved the model's ability to process events with high-frequency content like cymbal crashes, and the higher frame resolution was important for predicting events that repeated

rapidly, such as drum rolls. The resulting higher resolution network required more memory during training, so we also switched from processing batches of 20-second segments to 12-second segments.

For labels, we forced onset labels to occupy a single frame instead of being spread across 30ms of frames as they are in the original piano model. This also helped improve accuracy for rapidly repeating events. Finally, we added a 0.5 weight multiplier to the velocity loss to prioritize correct hit recognition during training.

We found that overfitting on the training data was a significant concern. The initial manifestation of this problem was that the trained model would transcribe only the first and last few seconds of an evaluation sequence. We suspect this was due to the bidirectional LSTM layer memorizing drum sequences that are simpler than the piano sequences this architecture was originally designed for (8 hits instead of 88 notes). Also, even though our training data has 35,217 audio examples due to our many drum kits, there are only 1,059 unique drum hit sequences.

To prevent overfitting, we used the standard techniques of reducing model capacity and adding dropout (Merity, Keskar, and Socher 2017). We decreased the size of the bidirectional LSTM layer from 128 to 64 units and added dropout at a rate of 50% to the outputs of the LSTM cells, but this alone was insufficient.

We also used a form of *mixup* (Zhang et al. 2017) for data augmentation and regularization. We created 500,000 training examples by randomly selecting pairs of examples from the training set, repeating the shorter of the examples until it was as long as the longer one, and then mixing their audio samples and underlying MIDI data together (prior to spectrogram or piano roll calculation) to form a new example, which is then split into 12-second chunks. This improved evaluation scores, but we still saw strongly divergent train/evaluation curves.

To create further diversity during training, we split those 500,000 examples into 1-second chunks. Then, at training time we spliced together random chunks into a 12-second example. We call this technique *Shuffled mixup* because it shuffles the order of many small chunks in addition to mixing examples together. This expands the *mixup* technique to sequence models and creates additional variety and better regularization during training.

With this final configuration, we no longer saw diverging train/evaluation curves. A comparison of these different techniques can be seen in Table 5.

After resolving the issue of overfitting to sequences, we also performed a coarse hyperparameter search and discovered that using a smaller convolutional stack prevented the model from overfitting to the particular characteristics of the drum sets in our training dataset. We reduced the number of filters in the convolutional layers from 32/32/64 to 16/16/32 and decreased the units in the fully connected layer from 512 to 256.

Our final model was trained with a batch size of 128 for 569,400 steps on 16 TPUv3 cores, which took about 3 days. We used the Adam optimizer with an initial learning rate of $1e-4$ and an exponential learning rate decay, reducing by

Model	Valid	Test	IDMT	ENST
Shuffled mixup	**88.71**	**83.40**	**85.72**	**76.89**
mixup	79.48	69.11	47.44	62.27
Unmodified	74.66	63.07	52.74	67.35

Table 5: Data augmentation and regularization ablation study. Results are F-measure scores calculated on E-GMD Validation, E-GMD Test, IDMT, and ENST. *Shuffled mixup* is the technique used when training our final OaF-Drums model. Training setup for the other methods is otherwise the same except that training was stopped after approximately 250k steps.

a factor of .98 every 10,000 steps. No early stopping strategy was used other than seeing that the train and evaluation curves had stabilized. We performed a coarse sequential search ($\approx$100 runs) over convolutional architectures, layer sizes, and input resolutions to arrive at the configuration used in the paper.

Code for training and evaluation along with a pre-trained model for inference is available on GitHub: https://goo.gl/magenta/onsets-frames-code.

5 Evaluation

Table 6 compares classifier scores for a variety of models and datasets. F-measure (also known as F1 score) is used as the evaluation metric, with a 50ms tolerance window of ground truth annotations for detected onsets as is consistent with the prior studies. We use the *mir_eval* package for metrics calculation (Raffel et al. 2014).

We compare against the two other models that were also used in the listening study. These models are ADTLib[3] and DrumTranscriptor[4] (DT), which are from Southall et al. (2017) and Vogl et al. (2018) respectively. ADTLib is trained on the standard 3 hit ADT task, while DrumTranscriptor is capable of transcribing 18 hits.

The public implementation of DrumTranscriptor is an ensemble 5 models trained on 5 different datasets: TMIDT, TMIDT balanced, ENST, MDB, and RBMA. We refer to this as DrumTranscriptor Ensemble (DT-Ensemble). This contrasts with the single DrumTranscriptor model (DT) in the paper, the best variant of which is trained only on TMIDT. We use DT-Ensemble for our listening study as it outperforms the DT model.

We train OaF-Drums on the E-GMD dataset and evaluate it on IDMT (3-hit standard) and ENST (multi-hit standard) for comparisons to other models.

IDMT Evaluation

IDMT was chosen primarily due to its consistent use in prior studies. It contains only the standard 3 hits (KD, SN, HH). In order to evaluate OaF-Drums in the simpler ADT task, we grouped the 7 possible drum hit predictions into the 3 hits. This grouping is shown in Table 2. This is somewhat different than other models we compare against that were trained

---

[3]https://github.com/CarlSouthall/ADTLib
[4]http://ifs.tuwien.ac.at/~vogl/dafx2018/

		F-measure				Listening Wins
Model	Training Dataset(s)	IDMT	ENST	E-GMD	E-GMD (vel)	Loop Loft
OaF-Drums	E-GMD	85.72	76.89	83.40	61.70	**919**
DT-Ensemble*	TMIDT(-Bal), MDB, ENST, RBMA	91.49	82.96	64.98	×	677
DT	TMIDT	×	68.00	×	×	×
ADTLib	ENST-3	83.12	×	×	×	372

Table 6: F-measures and listening study results from Section 6. Note the OaF-Drums model wins the listening study by a significant margin despite achieving comparable classification results to other models. The asterisk on DT-Ensemble* highlights that the model is actually an ensemble of 5 models trained on 5 different datasets. We use the DT-Ensemble in the listening study as it outperforms the single DT model. OaF-drums is the only model that predicts velocities, so it is the only model to be evaluated on E-GMD velocity labels. Since the various models are trained on different datasets, we compare classifier scores across a range of datasets, and perform the listener studies on the Loop Loft dataset, on which none of the models have been trained.

to predict only those 3 hits and ignore other audio events. We believe this comparison is reasonable because both training/evaluation methods incorporate *a priori* knowledge of what hits need to be predicted. This is yet another example of how different hit mapping strategies makes ADT evaluation difficult. Ultimately, we believe any comparison of models needs to incorporate a perceptual component as we do in the Listening Test in Section 6.

We evaluated against ADTLib and DT-Ensemble for IDMT. DT-Ensemble uses the same 7 hit grouping that OaF-Drums did. ADTLib only uses the 3 hit grouping and was trained on ENST only considering the standard 3 hits (ENST-3). The IDMT results for ADTLib, OaF-Drums and DT-Ensemble are shown in Table 6. All models perform rather well, with DT-Ensemble having the best score followed by OaF-Drums.

A full IDMT evaluation against the state of the art models reviewed in Wu et al. (2018) is in the Supplement. OaF-Drums has the 3rd best average F-measure of the 11 models. All the other models perform the standard 3 hit classification like ADTLib. The competitive score for OaF-Drums adds confidence that it performs well in the simpler ADT task, especially considering the model has been trained for more complex classification in the number of drum hits and added velocity prediction.

ENST Evaluation

We evaluate against ENST to compare our model in the multi-hit scenario, beyond the typical 3 hit ADT task. There are only a few models that attempt to model beyond 3 hits (Dittmar and Uhle 2004; Vogl, Widmer, and Knees 2018; Choi and Cho 2019; Cartwright 2018), and there is no standardization of evaluation for multi-hit models. There are also a very small number of public datasets that have multi-hit annotation, and within those datasets there is inconsistency in number and type of drum hits used.

Of the multi-hit models, Vogl et al. (2018) appear to have the best generalized performance across different datasets, and a public model implementation (DT-Ensemble) was available for additional inference for the listening study. Therefore, we elected to use that work as a proxy for the current state of the art in the multi-hit scenario.

Multi-hit comparison is a non-trivial task since DT-Ensemble is capable of classifying 18 different drum hits, which contrasts to the 25 different drum hits labeled in E-GMD, and the 20 different drum hits labeled in ENST. While there are some consistent mappings between drum hits in each domain, for example, KD, there is a lot of variation and ambiguity in mapping other categories such as cymbals and toms. We elected to evaluate the multi-hit task on a reduction of seven hits shown in Table 2. This seven-hit mapping is comparable to the eight-hit model of DT and DT-Ensemble because Clave (the eighth kind of hit) is not used in either our training or evaluation datasets. DT-Ensemble never predicted Clave during evaluation.
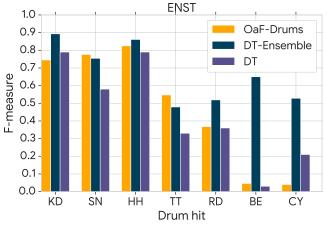
The F-measure results for ENST are shown in Table 6. OaF-Drums outperforms DT, but both are outperformed by DT-Ensemble, which is expected since DT-Ensemble is trained on ENST. The F-measure results broken down by drum hit are shown in Figure 2.

When broken down by hit, the F-measure results reveal stark contrasts in performance for different hits. Events such as Bells (BE) are rare and have significant variation between datasets, leading to poor generalization of models not trained on the dataset (OaF-Drums and DT for ENST, and DT-Ensemble for E-GMD).

Some attempts have been made to combat this behavior. Applying different weights to onsets in the loss function can help in some cases (Cartwright 2018; Vogl, Dorfer, and Knees 2017), but it does not appear effective in the cases of extremely sparse onsets. A more promising approach would be to re-balance the dataset to a more even distribution of onsets, which is explored with the TMIDT dataset in (Vogl, Widmer, and Knees 2018). The balanced dataset carried a trade-off in that model however, since per hit F-measures were much more even but overall F-measure notably decreased.

E-GMD Evaluation

As a final test, we evaluate OaF-Drums and DT-Ensemble against E-GMD. We elect to reduce all drum hit classes down to the same seven classes used in the ENST test as shown in Table 2. The results of E-GMD are shown in Table 6. The F-measure scores for both test and validation are
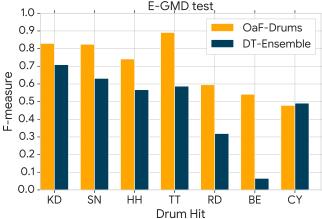
Figure 2: The F-measure results per hit on ENST and E-GMD test. The ordering of bars from left to right is OaF-Drums, DT-Ensemble, DT for ENST and OaF-Drums, DT for E-GMD test. DT-Ensemble included ENST in its training set while OaF-Drums and DrumTranscriptor did not. Events such as Bells (BE) are rare and have significant variation between datasets, leading to poor generalization of models not trained on the dataset (OaF-Drums and DT for ENST, and DT-Ensemble for E-GMD).

shown in the Supplement. Not surprisingly, OaF-Drums outperforms DT-Ensemble. While OaF-Drums did not train on any of the sequences in the E-GMD test subset, the training dataset did have audio from the same drum kits.

We also evaluate OaF-Drums performance using an F-measure score that includes velocity predictions as described in (Hawthorne et al. 2018). We only evaluate OaF-Drums on velocities, as the other models do not predict velocity labels. Results are again shown in Table 6. Results for both test and validation splits are shown in the Supplement.

Across all datasets, we see that OaF-Drums performs very competitively in an F-measure comparison. This is a good sign of generalization for the model, that it can consistently perform well across datasets not seen during training.

6 Listening Test

To measure the perceptual quality of our transcription model, we conducted a listening test where raters compared synthesized transcriptions to original recordings. We opted not to use any samples from the standard transcription datasets so that no model would have a particular advantage, and instead used 496 examples drawn from a commercial drum loop set (Loop Loft)[5]. Transcription model outputs were synthesized using FluidSynth[6] and the SGMv2.01-Sal-Guit-Bass-V1.3 SoundFont[7]. We also decided to focus on comparing models with 7 or fewer output classes because that made it clear how to define a consistent set of General MIDI instruments for synthesis. We mapped all model outputs to the following General MIDI instruments: 36 (Bass Drum 1), 38 (Acoustic Snare), 42 (Closed Hi Hat), 47 (Low-Mid Tom), 49 (Crash Cymbal 1), 51 (Ride Cymbal 1), 53 (Ride Bell).

Synthesizing model output like this has definite limitations. In particular, the drum kit in the SoundFont may sometimes sound very different from the original recording, and velocity changes in the SoundFont typically just scale the volume of the same sample without taking into account the changing physical response of a more or less forceful hit. However, the listening test has the significant advantage of allowing direct comparison of different models in the domain we care about (human perceptual audio similarity) using the same set of sounds.

We compare the outputs of ADTLib, DT-Ensemble, OaF-Drums, and OaF-Drums with output velocities fixed to a constant level. Only OaF-Drums outputs velocity predictions, all others used a fixed velocity of 100.

For each of the 496 examples, we selected a random 10-second clip (or the entire example if it was less than 10 seconds) and the associated synthesized outputs from each of the models. We then generated questions for each of the 6 possible pairwise comparisons between the models, resulting in a total of 2,976 questions. For each question, we asked raters which output best captured the content of the original clip and asked them to rate their choice on a 5-point Likert scale. Figure 3 shows the number of comparisons in which each source was preferred, with the OaF-Drums model having the overall highest number of wins.

Table 7 shows the results of comparing our model with and without velocity predictions and clearly demonstrates the perceptual importance of velocity.

Model	Number of wins
OaF-Drums w/ velocity	919
OaF-Drums w/o velocity	456

Table 7: Listening test results comparing output of the E-GMD 8 model with velocity predictions and with velocity fixed to a constant level.

---

[5]https://www.thelooploft.com/products/nate-smith-drums-bundle

[6]http://www.fluidsynth.org/

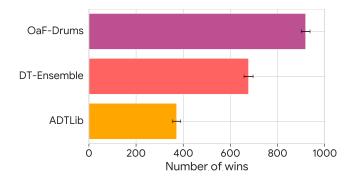[7]https://sites.google.com/site/soundfonts4u/

Figure 3: Results of our listening tests, showing the number of times each model won in a pairwise comparison. Black error bars indicate estimated standard deviation of means.

A Kruskal-Wallis H test of the ratings showed that there is at least one statistically significant difference between the models: $\chi^2(2) = 559.19, p < 0.001$ (7.0846e−121). A post-hoc analysis using the Wilcoxon signed-rank test with Bonferroni correction showed that there were statistically significant differences between all model pairs with $p < .001/6$.

The online supplement includes examples of listening test comparisons at https://goo.gl/magenta/e-gmd-examples.

7 Conclusion and Future Work

In this work we explored improving perceptual quality in ADT. We introduced the Expanded Groove MIDI Dataset and use the included velocity annotations to train an OaF-Drums model with added velocity predictions. Despite achieving similar results on classification metrics, we showed that multi-hit velocity prediction is well-aligned to the downstream task of generating audio, giving significant improvements in perceptual quality as determined by listening tests.

This work also highlights the value of listening studies in evaluating transcription systems, as an example of classifier outputs as inputs to generative systems. Incorporating such studies into the standard suite of classification metrics has the potential to expand the downstream applications of ADT and provide a fair comparison of models between different datasets and architectures.

Future work could include better representation of more drum hits and combining this model with a pitched automatic music transcription model for full music ensemble transcription.

References

Cartwright, M. 2018. Increasing Drum Transcription Vocabulary Using Data Synthesis. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*.

Choi, K.; and Cho, K. 2019. Deep Unsupervised Drum Transcription. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Delft, Netherland*.

Dittmar, C.; and Gärtner, D. 2014a. Real-Time Transcription and Separation of Drum Recordings Based on NMF Decomposition. In *DAFx*, 187–194.

Dittmar, C.; and Gärtner, D. 2014b. Real-Time Transcription and Separation of Drum Recordings Based on NMF Decomposition. In *DAFx*.

Dittmar, C.; and Uhle, C. 2004. Further steps towards drum transcription of polyphonic music. In *in Proc. 11th AES Conv*.

Gillet, O.; and Richard, G. 2006. ENST-Drums: an extensive audio-visual database for drum signals processing. In *Proc. Intl. Society for Music Information Retrieval Conf.*, 156–159. ISMIR.

Gillick, J.; Roberts, A.; Engel, J.; Eck, D.; and Bamman, D. 2019. Learning to Groove with Inverse Sequence Transformations. *arXiv e-prints* arXiv:1905.06118.

Hawthorne, C.; Elsen, E.; Song, J.; Roberts, A.; Simon, I.; Raffel, C.; Engel, J.; Oore, S.; and Eck, D. 2018. Onsets and frames: Dual-objective piano transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*.

Lindsay-Smith, H.; McDonald, S.; and Sandler, M. 2012. Drumkit transcription via convolutive NMF. In *International Conference on Digital Audio Effects (DAFx), York, UK*.

Merity, S.; Keskar, N. S.; and Socher, R. 2017. Regularizing and optimizing LSTM language models. *arXiv preprint arXiv:1708.02182* .

Miron, M.; Davies, M. E. P.; and Gouyon, F. 2013. An open-source drum transcription system for Pure Data and Max MSP. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* 221–225.

Miron, M.; Davies, M. E. P.; and Gouyon, F. 2013. An open-source drum transcription system for Pure Data and Max MSP. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 221–225. ISSN 2379-190X. doi:10.1109/ICASSP.2013.6637641.

Raffel, C.; McFee, B.; Humphrey, E. J.; Salamon, J.; Nieto, O.; Liang, D.; Ellis, D. P.; and Raffel, C. C. 2014. mir_eval: A transparent implementation of common MIR metrics. In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer.

Southall, C.; Stables, R.; and Hockman, J. 2017. Automatic Drum Transcription for Polyphonic Recordings Using Soft Attention Mechanisms and Convolutional Neural Networks. In *ISMIR*.

Southall, C.; Stables, R.; and Hockman, J. 2018a. Improving Peak-picking Using Multiple Time-step Loss Functions. In *19th International Society for Music Information Retrieval Conference*. ISMIR.

Southall, C.; Stables, R.; and Hockman, J. 2018b. Player Vs Transcriber: A Game Approach To Data Manipulation For Automatic Drum Transcription. In *ISMIR*, 58–65.

Southall, C.; Wu, C.-W.; Lerch, A.; and Hockman, J. 2017. MDB Drums: An annotated subset of MedleyDB for automatic drum transcription. In *Extended abstracts for the Late-Breaking Demo Session of the 18th International Society for Music Information Retrieval Conference*.

Stables, R.; Hockman, J.; and Southall, C. 2016. Automatic Drum Transcription using Bi-directional Recurrent Neural Networks. In *17th International Society for Music Information Retrieval Conference*.

Ueda, S.; Shibata, K.; Wada, Y.; Nishikimi, R.; Nakamura, E.; and Yoshii, K. 2019. Bayesian Drum Transcription Based on Nonnegative Matrix Factor Decomposition with a Deep Score Prior. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 456–460. IEEE.

Vogl, R.; Dorfer, M.; and Knees, P. 2016. Recurrent Neural Networks for Drum Transcription. In *ISMIR*, 730–736.

Vogl, R.; Dorfer, M.; and Knees, P. 2017. Drum transcription from polyphonic music with recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 201–205. IEEE.

Vogl, R.; Dorfer, M.; and Knees, P. 2017. Drum transcription from polyphonic music with recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 201–205. ISSN 2379-190X. doi:10.1109/ICASSP.2017.7952146.

Vogl, R.; Dorfer, M.; Widmer, G.; and Knees, P. 2017. Drum Transcription via Joint Beat and Drum Modeling Using Convolutional Recurrent Neural Networks. In *ISMIR*.

Vogl, R.; Widmer, G.; and Knees, P. 2018. Towards Multi-Instrument Drum Transcription. In *the 21st International Conference on Digital Audio Effects*. DAFx.

Von Neumann, J.; Morgenstern, O.; and Kuhn, H. W. 2007. *Theory of games and economic behavior (commemorative edition)*. Princeton university press.

Wu, C.-W.; Dittmar, C.; Southall, C.; Vogl, R.; Widmer, G.; Hockman, J.; Muller, M.; and Lerch, A. 2018. A Review of Automatic Drum Transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26(9): 1457–1483. ISSN 2329-9304. doi:10.1109/taslp.2018.2830113. URL http://dx.doi.org/10.1109/TASLP.2018.2830113.

Wu, C.-W.; and Lerch, A. 2015. Drum Transcription Using Partially Fixed Non-Negative Matrix Factorization with Template Adaptation. In *ISMIR*, 257–263.

Wu, C.-W.; and Lerch, A. 2018. From Labeled to Unlabeled Data – On the Data Challenge in Automatic Drum Transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 445–452. Paris, France: ISMIR. doi:10.5281/zenodo.1492447. URL https://doi.org/10.5281/zenodo.1492447.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* .

Supplement

Model	Avg	KD	SN	HH
NMFD	90.25	95.87	83.41	91.47
SANMF	86.53	96.40	71.70	91.50
OaF-Drums	85.72	90.21	78.82	84.87
GRUts	85.14	92.49	70.30	92.64
tanhB	84.69	96.69	69.38	87.99
lstmpB	83.12	96.16	70.24	82.95
PFNMF	83.02	94.78	76.13	78.15
RNN	80.92	88.82	61.14	92.78
ReLUts	80.54	91.47	58.97	91.29
AM1	79.69	95.91	81.16	62.00
AM2	79.48	92.45	78.35	67.63

Table 8: F-measure performance against IDMT, showing the average, and per-instrument performance. The table is sorted in order of best average F-measure performance. Scores for models other than OaF-Drums are from the "eval cross" experiment described in Wu et al. (2018).

Model	Validation	Test
OaF-Drums	88.71	83.40
DT-Ensemble	64.07	63.98

Table 9: F-measure performance against E-GMD validation and test.

Model	Validation (Velocity)	Test (Velocity)
OaF-Drums	64.97	61.70

Table 10: F-measure performance including velocity prediction accuracy against E-GMD validation and test. Only OaF-Drums scores are calculated because it is the only model that predicts velocity.
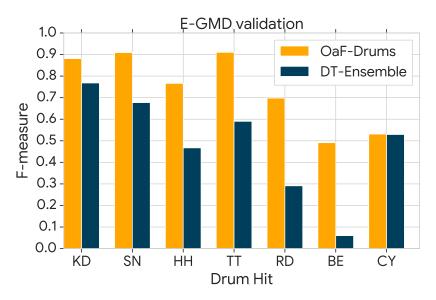
Figure 4: The F-measure results per hit on E-GMD validation splits. The ordering of bars from left is OaF-Drums, DT-Ensemble.

Layer	Size	Filters	Stride
Log Mel Spectrogram	250 bins		
Conv	16	3x3	1x1
BatchNorm			
Conv	16	3x3	1x1
BatchNorm			
MaxPool		1x2	1x2
Dropout		Keep 25%	
Conv	32	3x3	1x1
BatchNorm			
MaxPool		1x2	1x2
Dropout		Keep 25%	
Dense	256		
Dropout		Keep 50%	
Bidirectional LSTM	64		
LSTM Dropout		Keep 50%	
Dense	88		
Sigmoid Cross Entropy			

Table 11: Onset prediction architecture

Layer	Size	Filters	Stride
Log Mel Spectrogram	250 bins		
Conv	16	3x3	1x1
BatchNorm			
Conv	16	3x3	1x1
BatchNorm			
MaxPool		1x2	1x2
Dropout		Keep 25%	
Conv	32	3x3	1x1
BatchNorm			
MaxPool		1x2	1x2
Dropout		Keep 25%	
Dense	256		
Dropout		Keep 50%	
Dense	88		
Mean Squared Error			

Table 12: Velocity prediction architecture