TRAINABLE DATA MANIPULATION WITH UNOBSERVED INSTRUMENTS

Carl Southall, Ryan Stables, Jason Hockman

Digital Media Technology Lab
Birmingham City University
{carl.southall, ryan.stables, jason.hockman}@bcu.ac.uk

ABSTRACT

Machine learning algorithms are the core components in a wide range of intelligent music production systems. As training data for these tasks is relatively sparse, data augmentation is often used to generate additional training data by slightly altering existing training data. User-defined techniques require a long parameter tuning process and typically use a single set of global variables. To address this, a trainable data manipulation system, termed *player vs transcriber*, was proposed for the task of automatic drum transcription. This paper expands the player vs transcriber model by allowing unobserved instruments to also be manipulated within the data augmentation and sample addition stages. Results from two evaluations demonstrate that this improves performance and suggests that trainable data manipulation could benefit additional intelligent music production tasks.

1. INTRODUCTION

For a diverse range of intelligent music production related tasks—such as melody generation, automatic mixing and automatic drum transcription (ADT)—a large proportion of the state-of-the-art systems utilise machine learning algorithms. For these systems to perform as expected, they need to be trained using data that accurately represents the task. Although datasets do exist, they typically contain a relatively small number of examples compared to more mature research areas such as hand-written digit identification [1]. This results in gaps within the task representation in which system performance is greatly reduced. To overcome this limitation data augmentation is often used to generate new training examples by slightly altering pre-existing examples [2]. This results in a greater number of training examples without performing time-consuming manual annotation and often results in improved performance [3]. Data augmentation is typically performed using user-defined algorithms and settings [2]. This restricts the augmentation procedure as a long parameter process is required to achieve the best results. Additionally, as the parameter setting process is time consuming, the augmentation methods are usually determined with a single set of global variables. Thus they perform the same operation on all examples and do not take context variations into consideration. To overcome these restrictions within ADT, a new system, termed player vs transcriber (PvT) was proposed in [4]. Influenced by generative adversarial networks [5], PvT incorporates trainable data manipulation into a single end-to-end network which requires minimal parameter tuning. In an attempt to undermine the accuracy of a transcriber model (i.e., an existing state-of-the-art ADT system [6, 7]), a player model seeks to exploit poorly defined areas of the feature space through a manipulation of training data. This was achieved through learned data manipulation variables in the player network, which are used to define the manipulation coordinates of the transform. This enables the player model to manipulate data depending on its content instead of relying on a set of global variables and resulted in an increase in performance (up to 0.04 F-measure).

The remainder of this paper is structured as follows: In Section 2, the PvT model is extended to include additional unobserved instrumentation. An overview of the evaluation undertaken is provided in Section 3. Section 4 presents results and conclusions and future work are presented in Section 5.

2. INCLUDING UNOBSERVED INSTRUMENTS

In the original PvT paper, only samples of the observed drum instruments (i.e., kick drum (KD), snare drum (SD) and hi-hat (HH)) were included within the sample addition stage. However, in cases where the audio contains additional unobserved percussive and melodic instruments, these also have a significant contribution to the feature space. In this paper, the PvT model is expanded to identify whether including unobserved instruments within the sample addition stage can further increase performance.

2.1. Updating the Player Model

To include unobserved instruments within the PvT model only the player model is altered. As such, the other three stages—feature generation, the transcriber and peak-picking—remain the same [4]. Also, within the player model, only the sample addition stage is altered and so the data augmentation stage also stays the same; however, it is additionally performed on the unobserved samples. Figure 1 presents the updated player model which includes unobserved instrument samples. The same process is used as for the observed drum instrument samples; however, neither existing or output targets are required for the unobserved instruments. The unobserved instrument samples are added to the augmented existing spectrogram $X_{aug} \in \mathbb{R}^{T,F}$, after

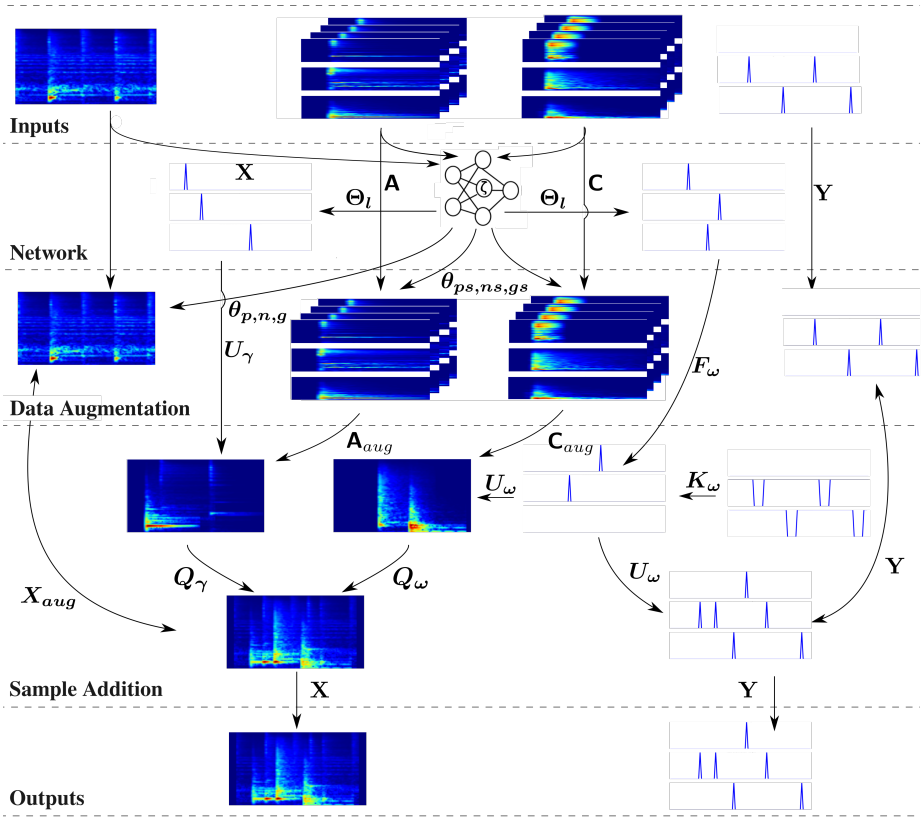

Figure 1: Overview of updated player model with two sets of samples (observed drums (middle left) and unobserved instruments (middle right)) utilised within the data augmentation and sample addition stages.

the observed drum instruments, using the same process as the original PvT model:

$$X = X_{aug} + Q_\gamma, \tag{1}$$

Here, γ is the unobserved sample number and Γ determines how many samples of each unobserved instrument are added. The new spectrogram $Q_\gamma \in \mathbb{R}^{T,F}$ is calculated using:

$$i_\gamma^\delta = \frac{\theta_{l,\gamma}^\psi}{max(\theta_{l,\gamma}^\psi)}, \tag{2}$$

$$f_\gamma^\delta = ReLU(i_\gamma^\delta + \epsilon - max(i_\gamma^\delta))\frac{1}{\epsilon}, \tag{3}$$

$$Z_\gamma^\delta = pad(A_{aug,\gamma}^\delta, T, T), \tag{4}$$

$$E_\gamma^{t,\delta} = Z_\gamma^{t+b,\delta} : Z_\gamma^{t+b+T,\delta}, \tag{5}$$

$$Q_\gamma = \sum_{\delta=1}^{\Delta}\sum_{t=1}^{T} e_\gamma^{t,\delta} f_\gamma^{t,\delta}, \tag{6}$$

which is a reduced version of the the observed drum instruments process with $F \in \mathbb{R}^{\Delta,T}$ being used in the last equations instead of U. δ is the instrument number (e.g., toms, cymbals), $A \in \mathbb{R}^{T,\Delta,F}$ are the unobserved instrument samples and $A_{aug} \in \mathbb{R}^{T,\Delta,F}$ are the augmented unobserved samples, which undergo the same process as C_{aug} [4].

3. UNOBSERVED PLAYER MODEL EVALUATIONS

To determine whether including unobserved instruments in the PvT model improves performance, two evaluations are performed using three datasets (ENST Drums [8], MDB Drums [9] and RBMA [10]), two evaluation strategies (random and subset [7]) and the F-measure metric. The first evaluation focuses on incorporating unobserved drum instruments (termed the DTP context [7]), and the second evaluation focuses on incorporating unobserved drum and melodic instruments (termed the DTM context). In both cases the unobserved `AAE` version of the PvT model is compared to the observed `AAE` version of the PvT model with comparable settings [4].

3.1. Unobserved Drum Instruments

The first evaluation incorporates different unobserved drum instruments within the PvT model. Seven possible combina-

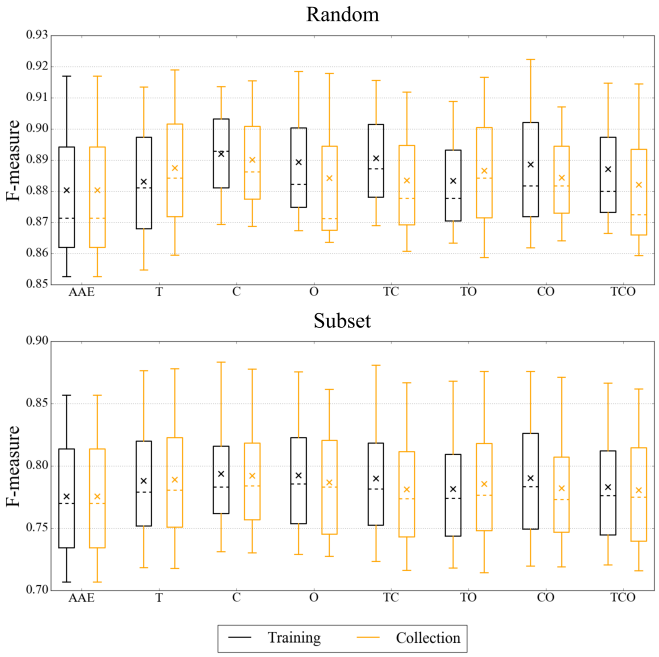

Figure 2: Results for the unobserved PvT system for the DTP context.

tions are produced from the three unobserved drum classes—toms (T), cymbals (C), and other percussion (O). For each case, the observed and unobserved total sample numbers (Ω and Z, respectively) are set to 2 and the max pooling layer size is altered to ensure that the total number of parameters of the player model is comparable to the transcriber model. A training and a collection set of unobserved drum samples are utilised; a training set is gathered from the isolated drum sample files contained within the ENST Drums dataset and the collection set is obtained from online resources. In total, the training set is comprised of 149 tom, 134 cymbal, and 43 other percussion samples and the collection set is comprised of 50 tom, 58 cymbal and 66 other percussion samples. All other settings are the same as the original PvT model.

3.2. Unobserved Drum and Melodic Instruments

The second evaluation incorporates different unobserved drum and melodic instruments within the PvT model. The five groups of instruments used in [11] are utilised (non-pitched percussion (N), pitched percussion (P), wind instruments (W), bowed strings (B), and vocals (V)), resulting in 30 different configurations. As none of the datasets contain melodic instrument samples, a collection set is created using only the datasets from [11] and online resources. In total, there are 800 non-pitched percussion, 3874 pitched percussion, 801 wind instrument, 1270 bowed string, and 310 vocal onsets.

4. RESULTS

4.1. Unobserved Drum Instruments

Figure 2 presents the DTP context results for the unobserved PvT model. The top diagram presents the F-measure results for the random evaluation strategy and the bottom diagram presents the results for the subset evaluation strategy. The crosses ('x') represent the mean, the dashed lines ('-') depict the median, and the box plots show the range across folds. The black box plots present systems that utilise training samples and the orange box plots depict collection samples. As the observed PvT model does not utilise any unobserved drum samples then the results are the same. For both training strategies, systems that utilise unobserved drum samples achieve higher mean and median F-measures than the original PvT system (`AAE`). This highlights that it is indeed beneficial to enable the system to both augment and place unobserved drum samples. Utilising multiple unobserved drum instruments does not achieve higher results than using a single instrument. This could be due to the system being able to create unrealistic situations where a large number of instruments are overlapping. As witnessed for the observed PvT system, higher improvements are achieved for the subset strategy than the random strategy; however, this gap is smaller than for the observed PvT system. As with the original PvT system, including unobserved drum instruments improves the generalisability of the systems. T-tests performed across folds and tracks highlight that the improvement achieved by all combinations is significant for both training strategies ($\rho < 0.05$). For all but three of the combinations, utilising samples from the training data achieves higher performance than using samples collected from random sources. This suggests that the system can learn the specific timbral features and instrument interactions.

4.2. Unobserved Drum and Melodic Instruments

Figure 3 presents the DTM results for the unonbserved PvT system using the same statistical inspection as in the DTP context. In both training strategies utilising the unobserved samples within the player model improves the performance of the system. The fact that the increase in performance in the subset evaluation is higher and more frequent than the random strategy, reinforces that the PvT model improves the generalisability of the systems as the highest performances are achieved by utilising the additional unobserved samples. However, unlike for the DTP context, not all of the combinations achieve a higher mean F-measure than the observed PvT system (`AAE`). This observation demonstrates that some groups improve and some groups hinder performance, with non-pitched percussion and vocal samples resulting in the greatest performance improvement. The reduction in performance observed for some of the combinations could be explained by the diversity of samples. For example, there are instruments that are not contained within the audio files

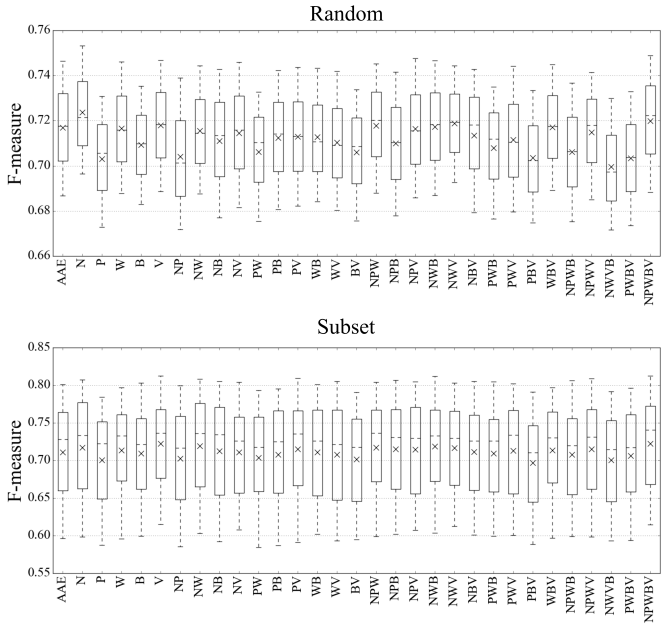

Figure 3: Results for the unobserved PvT system for the DTM context.

used. Although this results in the system generalising to a wide range of instruments, it reduces how effective the system is at an individual instrument level. Also, the smaller increase in performance, compared to the DTP context, can be explained by the complexity and diversity of the DTM context (i.e., a much larger range of possibilities). In both strategies the highest results were achieved by incorporating all of the samples (NPWBS). This finding contradicts that of the DTP context and suggests that the best performance is achieved by enabling the PvT model to represent as much of the feature space as possible. Results from t-tests highlight that all of the improvements witnessed in random and the improvements of N, V, NW, PV, NPW, NWV, NBV, WBV and NPWBV in subset are significant ($\rho < 0.05$).

5. CONCLUSIONS

The main findings from incorporating unobserved percussion and melodic instrument samples within the PvT model, is that enabling the system to augment and add new samples of unobserved instruments results in substantial improvements. This demonstrates that it is beneficial to incorporate unobserved instances into a trainable data manipulation process. Possible future work could be to turn the PvT model into a more generalised and modular framework which will enable trainable data manipulation to be easily applied to a range of intelligent music production tasks.

6. REFERENCES

[1] Y. LeCun, "The MNIST database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[2] B. McFee, E. J. Humphrey, and J. P. Bello, "A software framework for musical data augmentation," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, (Malaga, Spain), pp. 248–254, 2015.

[3] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[4] C. Southall, R. Stables, and J. Hockman, "Player vs transcriber: A game approach to data manipulation for automatic drum transcription," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, (Paris, France), pp. 58–65, 2018.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[6] C. Southall, R. Stables, and J. Hockman, "Improving peak-picking using multiple time step loss functions," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, (Paris, France), pp. 313–320, 2018.

[7] C.-W. Wu, C. Dittmar, C. Southall, R. Vogl, G. Widmer, J. Hockman, M. Müller, and A. Lerch, "A Review of Automatic Drum Transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1457–1483, 2018.

[8] O. Gillet and G. Richard, "ENST-drums: An extensive audio-visual database for drum signals processing," in *Proceedings of the 7th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 156–159, 2006.

[9] C. Southall, C.-W. Wu, A. Lerch, and J. Hockman, "MDB drums an annotated subset of medleyDB for automatic drum transcription," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, (Suzhou, China), 2017.

[10] R. Vogl, M. Dorfer, G. Widmer, and P. Knees, "Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks.," in *ISMIR*, pp. 150–157, 2017.

[11] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, (Porto, Portugal), pp. 49–54, 2012.