AUTOMATIC DRUM TRANSCRIPTION USING THE STUDENT-TEACHER LEARNING PARADIGM WITH UNLABELED MUSIC DATA

Chih-Wei Wu, Alexander Lerch
Georgia Institute of Technology, Center for Music Technology
{cwu307, alexander.lerch}@gatech.edu

ABSTRACT

Automatic drum transcription is a sub-task of automatic music transcription that converts drum-related audio events into musical notation. While noticeable progress has been made in the past by combining pattern recognition methods with audio signal processing techniques, the major limitation of many state-of-the-art systems still originates from the difficulty of obtaining a meaningful amount of annotated data to support the data-driven algorithms. In this work, we address the challenge of insufficiently labeled data by exploring the possibility of utilizing unlabeled music data from online resources. Specifically, a student neural network is trained using the labels generated from multiple teacher systems. The performance of the model is evaluated on a publicly available dataset. The results show the general viability of using unlabeled music data to improve the performance of drum transcription systems.

1. INTRODUCTION

Data availability, listed by Schedl et al. as one of the open challenges in the field of Music Information Retrieval (MIR) [21], is an important problem that concerns a large variety of data-driven MIR systems. To create intelligent music (analysis) systems, music data with detailed annotations is crucial as training input for machine learning algorithms. However, multiple constraints impede the availability of large datasets, including (i) the complexity and variety of music in terms of genres, instrumentation, tonality, etc., (ii) the difficult and time-consuming process of manually adding annotations which —- for most tasks — might also depend on perception and thus require multiple annotators, and (iii) intellectual property laws, restricting the compilation and sharing of music datasets. Many laudable efforts have been made to address (some of) these problems, leading to the release of new datasets or the extension of existing datasets. Nevertheless, the majority of the commonly used datasets for various MIR tasks is still limited in different aspects, which can impact research focus. For example, Benetos et al. pointed out that a large subset of Automatic Music Transcription (AMT) approaches only performed experiments on piano data for which the audio aligned ground truth was easily obtained [1]. This emphasis on piano may lead to models that are strongly biased towards piano-like instruments and cannot be generalized to other melodic instruments.

Automatic Drum Transcription (ADT), a sub-task in AMT that involves the extraction of drum events from audio signals, is also confined to the scope of the existing labeled datasets. Wu observed [30] that most of the ADT related datasets focus on collecting recordings of single drum hits [18, 24] and simple drum sequences without accompaniment [5]. Although these datasets provide the essential ingredients for building basic ADT systems, they cannot properly represent the real-world scenario of drum sounds embedded in a continuous stream of polyphonic audio sources. Thus, they might fail in addressing real-world use cases. The ENST drum dataset [8] partly compensates these drawbacks by offering more realistic and complex drum sequences with accompaniments, however, its size and diversity of music styles are still limited. Previous studies attempt to alleviate these issues through data augmentation [26, 30], but the inherent limitations of the datasets continue to impede the advancement of ADT systems.

One potential solution to addressing this challenge in a scalable way without introducing the additional cost of manual annotations is to explore the usefulness of the vast collection of unlabeled music data; this can be formulated as a *Semi-supervised Learning* problem as defined in the field of machine learning [3]. The general goal of this type of problem is to find the optimal solution given both labeled and unlabeled examples, and it has been applied successfully to different applications such as music genre classification [19], music genre tagging [13], and music emotion recognition [28].

Inspired by the above-mentioned approaches, this paper aims to address the issue of data availability in ADT systems by harnessing the information from the unlabeled music data. Specifically, this paper focuses on improving ADT performance on polyphonic mixtures. The contributions of this paper include: (i) new insights into the viability of using unlabeled music data in ADT tasks, (ii) a general scheme for integrating unlabeled data into ADT and other MIR systems, and (iii) the demonstration of potential improvements of ADT systems using the proposed method. The remainder of the paper is structured as follows: Sect. 2 provides an overview of ADT research and the student-teacher learning paradigm. In Sect. 3, we introduce our approach; the results and discussion are presented in Sect. 4. Sect. 5 provides a summary, conclusion, and directions of future work.

2. RELATED WORK

In the broadest definition of ADT, it can be described as the process of converting drum related audio events, such as drum onset times and playing techniques, into musical representations such as a score or sheet music. To simplify this task while still capturing the essence, most of the existing systems mainly focus on detecting the onset times of *Hi-Hat* (HH), *Snare Drum* (SD) and *Bass Drum* (BD). In many of the early systems, which are summarized by FitzGerald and Paulus [6], the focus was on transcribing signals containing only drum sounds.

Gillet and Richard propose to categorize automatic drum transcription systems into three categories [9]: (i) *segment and classify* [7, 9], which follows the basic pattern recognition approach by segmenting the signals into individual instances, and subsequently classifying each instance with pre-trained classifiers, (ii) *separate and detect* [5, 20, 29], in which the signal is converted into separated activation functions that represent the activities of different drums, followed by a simple peak picking process to identify their corresponding onset times, and (iii) *match and adapt* [31], which identifies the drum events by template matching using a set of pre-trained drum templates and customized distance measures; the templates are iteratively adapted throughout the process. In addition to these three categories, a language-model-based approach using Hidden Markov Models (HMM) [17] and a pattern-matching approach using bar information [23] have also been applied to ADT tasks in previous work.

Following the recent success in deep learning [10], several state-of-the-art ADT systems utilize Deep Neural Networks (DNNs). Specifically, Recurrent Neural Networks (RNNs), a DNN variant modeling the temporal dependency of the input using recurrently connected nodes, have been adopted for this task [22, 25, 26]. Although this method is capable of learning complicated representations of drums from the audio signals, it is extremely demanding in terms of the required amount of training data and computing power. To reach their full potential, DNNs require large amounts of training data; the sizes of currently available datasets appear to be insufficient, as exemplified by the performance degradation in polyphonic mixtures reported in several ADT systems [22, 26, 29]

To overcome the problem of possibly insufficient input data for data-hungry approaches such as DNNs, the idea of utilizing the unlabeled data seems very appealing. Recently, the concept of the student-teacher learning paradigm has emerged as an interesting way of incorporating unlabeled data in the training of DNNs. Originally proposed as a model compression method [2], the basic idea of student-teacher learning is to transfer the knowledge of a large teacher model into a small and concise student model with minimum performance loss; this process, referred by Hinton et al. as ”knowledge distillation” [11], is achieved by training the student model with the soft targets generated from the teacher model. In other words, instead of learning from the hard targets (i.e., the ground truth), the student model indirectly acquires the knowledge by mimicking the output from the teacher model. As demonstrated by Li et al. [16], this process can use labeled as well as unlabeled data. Successful applications of this paradigm can be found in tasks such as speech recognition [27] and multilingual models [4], in which superior performances from the student model have also been reported.

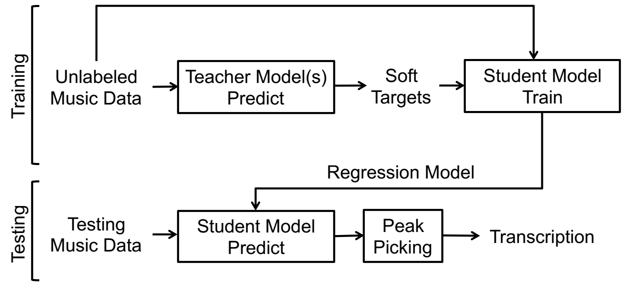

Figure 1. The flowchart of the proposed method

3. METHOD

3.1 System Overview

The processing steps of the proposed method, as shown in Figure 1, can be split into two phases, namely the training and testing phase. In the training phase, the unlabeled music data are passed through the teacher models in order to generate the soft targets. Specifically, these teacher models are ADT systems that will convert the audio signals into drum-related activation functions (i.e., soft targets). The same unlabeled music data and the generated soft targets will then be used to train a student model, which is a regression model that minimizes the differences between its output and the soft targets. In the testing phase, the trained student model predicts the drum activations of the test music data. Finally, a simple peak picking algorithm with an adaptive threshold will be used to identify the drum onset times from each activation function, producing the final transcription output. More elaborate descriptions of the teacher and student models can be found in the following sections.

3.2 Teacher Model

The teacher model used in this paper is the drum transcription system presented by Wu and Lerch [29]. This NMF-based ADT system is chosen for its simplicity, its lack of need for substantial amounts of training data, as well as the adaptability in polyphonic mixtures; it extends the basic NMF model to Partially-Fixed Non-negative Matrix Factorization (PFNMF) by assuming the co-existence of both percussive and harmonic components in the audio signals. More specifically, the template matrix is split into a predefined part containing the drum templates which kept fixed and not iteratively updated and a randomly initialized part

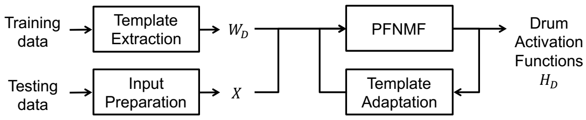

Figure 2. The flowchart of PFNMF [29]

for modeling the remaining harmonic components in the signal. Formally, this can be expressed as

$$X \approx W_{\mathrm{D}} H_{\mathrm{D}} + W_{\mathrm{H}} H_{\mathrm{H}}, \tag{1}$$

with X being a $m \times n$ magnitude spectrogram matrix with m frequency bins and n blocks, W_{D} and W_{H} representing the drum and harmonic dictionary matrices with a dimensionality of $m \times r_{\mathrm{D}}$ and $m \times r_{\mathrm{H}}$, and H_{D} and H_{H} their corresponding activation matrices with dimensionality of $r_{\mathrm{D}} \times n$ and $r_{\mathrm{H}} \times n$, respectively. r_{D} usually corresponds to the number of drums to detect (e.g., $r_{\mathrm{D}} = 3$ for the detection of HH, BD, and SD), and r_{H} is an user-defined parameter that varies according to the complexity of the target signal.

The basic flowchart of PFNMF is shown in Figure 2. It firstly decomposes the magnitude spectrogram of the polyphonic mixtures with a fixed pre-trained drum dictionary W_{D} and a randomly initialized harmonic dictionary W_{H}. Once the signal is decomposed, the NMF based activation function $H_{\mathrm{D}}(r,:)$ of each individual drum can be extracted, in which $r = \{1, 2, 3\}$ is the instrument index that corresponds to HH, BD, and SD, respectively. These activation functions can be interpreted as the activity level of each instrument over time, and a sharp peak indicates the presence of a single drum hit.

The conversion of the resulting activation functions into the soft targets takes another step of standard min-max scaling across the training data for each instrument; this process scales the soft targets to a numerical range between 0 and 1 and ensures the compatibility between the soft targets and the student model output (see Sect. 3.3). Finally, to introduce diversity into the soft targets, two PFNMF systems are created by initializing the algorithm with two different sets of drum dictionaries, forming an ensemble-like scenario that could potentially lead to better student performance.

3.3 Student Model

The proposed student model is a fully connected, feed-forward DNN with three hidden layers. A neural network is a graphical model that comprises multiple layers of interconnected non-linear units (i.e., neurons). The basic formulation of a neuron can be expressed in Eq. (2)

$$a_k^l = g\left(\sum_{j=1}^{M} W_j a_j^{l-1} + b_j^{l-1}\right), \tag{2}$$

in which a is the activation of the neuron, W is the weight matrix, b is the bias matrix, l is the layer index, j is the index of input neuron, and k is the index of output neuron; $g()$ is usually a non-linear function such as a *sigmoid*, *tanh* or *relu*. When multiple layers of neurons are stacked, the model creates a non-linear transformation from the input to the output, which allows the model to approximate any arbitrary function with great flexibility.

The architecture of the DNN in this paper is as follows: the input layer contains 1025 neurons that correspond to the size of the input representation. The first hidden layer comprises of 1025 neurons of *tanh* units with Batch Normalization [12]. The second and third hidden layers have 512 and 32 neurons with *relu* units, respectively. Finally, the output layer consists of 3 neurons with *sigmoid* units that represent the activities of three different drums (i.e., HH, SD, and BD). The architecture and type of neurons are selected based on the results of smaller-scale preliminary experiments, and the fully connected layers are chosen for their simplicity and generality. To solve the optimization problem of learning the weights W in a DNN, a stochastic gradient descent based optimization method, Adam [14], is selected as the optimizer. The student neural network is configured as a regressor that minimizes the mean squared error between its output and the soft targets. A mini-batch consisting of 640 instances is used for training, and the early stopping technique is applied to stop the training process when the loss decrease is less than 10^{-6} for three consecutive epochs.

3.4 Implementation

The input representation to both the teacher and student models is the magnitude spectrogram of the Short Time Fourier Transform (STFT) computed using a block size of 2048 and hop size of 512 samples with a Hann window applied to each block. Prior to the calculation of STFT, the audio signals are down-mixed to mono and resampled to a sampling rate of 44.1 kHz. The resulting magnitude spectrogram is a $m \times n$ matrix, in which $m = 1025$ and n equals the number of blocks.

For PFNMF, the authors' open source Matlab implementation [1] is used in our experiments. Since both the unlabeled music data and the test data are polyphonic mixtures, the harmonic rank r_{H} for the PFNMF is set to 50 as suggested [29]. To speed up the process, template adaptation is deactivated. The extraction of the pre-defined (fixed) drum templates takes place on two publicly available drum datasets, namely the SMT-DRUM dataset [5] and 200 drum machines. [2]

Preliminary experiments show that these two sets of templates exhibit capabilities of capturing different types of drum sounds, thus adding diversity to this learning paradigm. The construction of the drum dictionary involves the concatenation of all the spectra and the extraction of the median spectrum for each individual instrument. It should be noted that, since the ENST drum dataset is the main test dataset for evaluation, no single drum hits from ENST are

[1] https://github.com/cwu307/NmfDrumToolbox Last accessed: 2017/04/26

[2] http://www.hexawe.net/mess/200.Drum.Machines Last accessed: 2017/04/26

Experiments				Averaged F-measure		
Role		Method	# Training Data	HH	BD	SD
Teacher	Baseline	PFNMF (SMT)	N/A	0.69	0.80	**0.50**
Teacher	Baseline	PFNMF (200D)	N/A	0.68	0.85	0.48
	Baseline	PFNMF (SMT + 200D)	N/A	0.69	0.83	0.48
Student	Baseline	Linear SGD Regressor	200 * 4 = 800	0.43	0.69	0.43
Student	Proposed	DNN	200 * 4 = 800	**0.78**	**0.86**	0.45

Table 1. A comparison of the averaged F-measures between the proposed method and the baseline methods

used for template extraction in order to ensure the generality of the proposed approach.

The DNN is implemented in Python using Keras [3] with the Tensorflow [4] backend. The parameters of the optimizer are set to default.

To get the final transcription results for evaluation, a standard peak picking method with a signal adaptive median threshold is used [15]. The median threshold $t(n)$ can be computed using Eq. (3):

$$t(n) = \lambda * max(x) + median(x(n), p), \qquad (3)$$

in which x is a vector of novelty function, λ is the offset coefficient relative to the maximum value, p is the order (length) of the median filter, and the n is the block index. All systems are using the peak picking parameters $p = 0.1$ s and $\lambda = 0.12$ as described in [29]. No grid search is performed.

4. EXPERIMENTS

4.1 Dataset Description

The collection of the unlabeled data is a crucial step for ensuring a successful learning process. Generally speaking, the unlabeled dataset should have following attributes: (i) the collection should contain drums whenever possible, (ii) the collection should be diverse in terms of music genres or playing styles, (iii) the collection should contain no duplicates, and (iv) the collection should be as consistent as possible in terms of audio quality. To build a collection that meets the above-mentioned criteria, we compile a list from the Billboard Charts. [5] In particular, we start with an uniform distribution across a set of 4 genres selected for commonly featuring strong drum beats or rhythmic patterns, namely R&B/HipHop, Pop, Rock, and Latin. For this study, 200 songs from each genre has been selected. All the songs are cross-checked for duplicates, and a final list of 800 songs has been compiled and retrieved from Youtube [6] using open source Python library pafy. [7]

All songs are converted into mp3 files with a sampling rate of 44.1 kHz using ffmepg. [8] The source code for constructing the unlabeled music dataset is available online on Github. [9] In order to speed up the process while retaining diversity, only a segment of 30 s from each song is used for training. This segment starts at 30 s into the song in order to avoid possible inactivity at the beginning. Since the same unlabeled data is trained twice with two different sets of soft targets generated from two different teachers, the total duration of the training audio is 800 mins (approximately 13.5 hours), which is significantly larger than any existing drum dataset.

The most popular labeled drum dataset, ENST drum [8], is used as the test set for evaluation. This dataset consists of recordings from three different drummers performing on their own drum kits. The recordings from each drummer contain individual hits, short phrases of drum beats, drum solos, and short excerpts played with accompaniments. Since this paper focuses on ADT in polyphonic mixtures of music, only the *minus one subset* is used for evaluation. This subset has 64 tracks of polyphonic music with a sampling rate of 44.1 kHz. Each track in this subset has a length of approximately 50–70 s with a variety of playing styles. More specifically, the subset contains various drum playing techniques such as ghost notes, flam, and drag, which is close to a real-world setting [30]. The accompaniments are mixed with their corresponding drum tracks using a scaling factor of 1/3 and 2/3 in order to be consistent with prior studies [17, 22, 29]. Only the wet mix recordings of the dataset are used.

4.2 Experiment Setup

The performance of the following systems is evaluated and compared:

(i) PFNMF (SMT): a PFNMF system initialized with a drum dictionary matrix extracted from SMT-DRUM dataset. This baseline system is used as a teacher model to generate the soft targets

(ii) PFNMF (200D): a PFNMF system initialized with a drum dictionary matrix extracted from 200 drum machines dataset. This baseline system is the second teacher model for generating the soft targets

(iii) PFNMF (SMT + 200D): another baseline system by simply taking the averaged activation functions of the above systems as the prediction output

(iv) Linear SGD Regressor: a baseline student model using a simple linear regression with stochastic gradient descent optimization. A Python implementation

[3] https://keras.io Last accessed: 2017/04/27

[4] https://www.tensorflow.org Last accessed: 2017/04/27

[5] http://www.billboard.com/charts Last accessed: 2017/04/25

[6] https://www.youtube.com Last accessed: 2017/04/25

[7] https://pypi.python.org/pypi/pafy Last accessed: 2017/04/25

[8] https://ffmpeg.org/download.html Last accessed: 2017/04/25

[9] https://github.com/cwu307/unlabeledDrumDataset

Experiments				Averaged F-measure		
Role	Method	Genres	# Training Data	HH	BD	SD
Student	DNN	Rock	200 * 1 = 200	0.76	0.83	0.44
Student	DNN	Pop	200 * 1 = 200	**0.78**	**0.85**	0.45
Student	DNN	RnB	200 * 1 = 200	0.74	0.83	**0.48**
Student	DNN	Latin	200 * 1 = 200	**0.78**	0.83	0.44
Student	DNN	All	50 * 4 = 200	0.77	**0.85**	0.45

Table 2. A comparison of different student models trained with unlabeled music data of different genres

of this method from the open source library scikit-learn [10] is used with all parameters set to default values.

(v) DNN: the proposed student model

4.3 Metrics

The evaluation metrics follow the standard calculation of the precision (P), recall (R), and F-measure (F). To be consistent with [9, 22, 29], an onset is considered to be a match with the ground truth if the time deviation between reference and detected onset time is less or equal to 50 ms. It should be noted that some authors use more restrictive settings, compare, for instance, the 30 ms and 20 ms tolerance windows as used in [17] and [26], respectively.

4.4 Results

The experiment results are shown in Table 1. The reported accuracies are the averaged F-measures across all 64 tracks from the ENST minus-one subset. Since the proposed method does not use the ENST drum dataset for training purposes, a three-fold cross validation scheme as reported in [17, 22, 25, 26, 29] is not necessary; this ensures the generality of the proposed method, but prohibits the direct comparison of the results with other publications.

The evaluation results show that both teacher systems PFNMF (SMT) and PFNMF (200D) perform similarly except for BD. This could be due to the discrepancy of the pre-defined drum dictionaries. The 3rd simple baseline system PFNMF (SMT+200D) averaging the teacher outputs gives almost identical performance as the teacher systems. This result shows that a simple combination of the two teacher systems does not result in any improvement. This means either that the performance cannot be improved given the teacher information or that a more sophisticated method is required for combining the outputs. The student baseline system is a simple linear regression model trained using the student-teacher learning paradigm as described in Sect. 3. This baseline serves as a sanity check for the necessity of a complex model such as DNN. As expected, the performance of the linear regression model is the worst among all the evaluated systems, indicating the need of deploying a non-linear model in order to benefit from this training scheme. Finally, the proposed DNN-based student model is actually able to outperform both teachers with higher F-measures for both HH and BD. The results for the SD are somewhat inconclusive; here, one teacher outperforms all other systems. This could imply the similarity between the SD sounds in SMT and ENST dataset, but the inferior performance from the student model still needs further investigation.

Based on these results, another interesting question arises: does music genre play a role in the preparation of unlabeled data? To answer this question, a follow-up experiment has been conducted by training the DNN model with unlabeled data of each individual genre. The experiment results are shown in Table 2. In this experiment, the number of training samples is fixed at 200 in order to eliminate the influence of data size. For the *All* case, 50 songs from each genre are randomly selected. Interestingly, the best performance of different instruments, as highlighted in the table, belongs to different genres. This implies the advantage of having various genres in the training data, for they could potentially complement each other and boost the performance of the student model.

Although the cross-genre model trained on the equally distributed data does not achieve the highest accuracy in every individual instrument, it is still better than majority of the single-genre models and generally well-balanced. Overall, providing diverse unlabeled training data in terms of music genre seems to be beneficial in this learning paradigm.

From all of the above experiment results, the results for HH show the most obvious and consistent improvement over the teacher models. This observation leads to another question: where do these improvements come from? A closer look at the experiment results reveals the strength of the DNN student model. As shown in Table 3, the DNN student model outperforms the teacher models on both precision and recall for HH. The DNN student model also achieves the highest BD precision. Since these improvements in precision are achieved without sacrificing recall, they suggest a reduction in false positives from the student model output. One possible explanation is that the songs presented in the unlabeled music data have a higher agreement on HH sound; this allows the student model to acquire a more consistent internal representation of HH that leads to a more accurate estimation during testing.

It is noticeable that the DNN student model seems to consistently have problems detecting SD. Since the snare drum tends to have larger spectral overlap with the other instruments, it is conceivable that DNN student model will have difficulties learning a robust internal representation for this instrument. A collection of unlabeled data with a stronger presence of snare drum might be possibly able

[10] http://scikit-learn.org Last accessed: 2017/04/25

Method	HH		BD		SD	
	P	R	P	R	P	R
PFNMF (SMT)	0.77	0.69	0.74	**0.91**	**0.67**	**0.49**
PFNMF (200D)	0.75	0.68	0.82	0.90	0.60	**0.49**
DNN	**0.87**	**0.72**	**0.83**	0.89	0.60	0.44

Table 3. A comparison of precision (P) and recall (R) between student and teacher models

to alleviate the problem, however, this issue requires further investigation before any conclusion can be drawn. In general, this deficiency in SD is also consistent with the previous studies [17, 22, 25, 29], where the detection of Snare Drum in polyphonic mixtures has been reported as the most difficult task in ADT. It is also possible that the Snare Drum is for some reason particularly hard to detect in the ENST set that is commonly used for evaluation.

5. CONCLUSION

This paper presents a system for Automatic Drum Transcription based on the student-teacher learning paradigm with the unlabeled music data. The proposed method integrates two NMF-based ADT teacher systems with a DNN-based student model by transferring knowledge using unlabeled music data, and the evaluation results indicate the possibility of obtaining a student model that outperforms the teacher model based on this approach. This result is generally encouraging and demonstrates the great potential of using unlabeled music data in ADT tasks. The experiment results also imply the benefit of having relevant music genres in the unlabeled training data, which could lead to the construction of an improved unlabeled dataset in the future studies. The proposed method has the following advantages: first, the approach allows for complete separation between training and test data, therefore reducing the likelihood of over-fitting and supporting the claim of generality of this approach. Second, the proposed method is able to support data-driven approaches with the need of large amounts of training data given the availability of existing teacher models. Third, the proposed method could not only be easily applied to other ADT systems but also inform data-hungry systems from other transcription tasks or MIR problems in general. Last but not least, this learning scheme has the potential of summarizing multiple complicated teacher systems, providing competitive performance with one concise student model.

The possible future directions of this work are:

(i) *Increasing the number and diversity of teacher systems.* Since the proposed training scheme does not tie to any particular ADT approach, the teacher models can be easily swapped with other ADT expert systems. Intuitively, more teacher models should lead to a more versatile student model. However, the influence of having a more diverse pool of teacher systems still requires further investigation.

(ii) Varying architectures and approaches of the student models. In addition to DNNs, other neural networks architecture may have great potential of achieving better student performance as well. For instance, the RNN based model that incorporates the temporal information could be a good fit in the context of ADT tasks.

(iii) *Evaluating different input representations.* As reported by Cui et al. [4], the student model is able to outperform the teacher model especially when it is trained on the same soft targets but with a stronger input representation. Following this observation, one possible future direction of this work is to investigate the effectiveness of other input representations, such as CQT, Cepstrum, or Wavelet transforms.

(iv) *Evaluating alternative approaches for using unlabeled data.* To fully benefit from the unlabeled data, it is also worth investigating how the proposed method compares to other approaches such as unsupervised feature learning [19].

The presented work represents only a preliminary study of what the authors see as a likely path for the future of training MIR systems as the issue of an insufficient amount of annotated data is likely to get worse with increasing complexity of machine learning systems applied to MIR tasks. Drawing on the vast potential of using existing state-of-the-art MIR-systems as teachers and the overwhelming public availability of unlabeled music data might enable exciting ways of creating new and more powerful MIR systems.

6. REFERENCES

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

[2] Cristian Bucilu, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proc. of the International Conference on Knowledge Discovery and Data mining (SIGKDD)*, page 535, 2006.

[3] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 2006.

[4] Jia Cui, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, Tom Sercu, Kartik Audhkhasi, Abhinav Sethy, Markus Nussbaum-Thom, and Andrew Rosenberg. Knowledge Distillation Across Ensembles of Multiplingual Models for Low-resource Languages. In *Proc. of the International Conference on Acoustics, Speech*

and Signal Processing (ICASSP), pages 4825–4829, 2017.

[5] Christian Dittmar and Daniel Gärtner. Real-time Transcription and Separation of Drum Recording Based on NMF Decomposition. In *Proc. of the International Conference on Digital Audio Effects (DAFX)*, pages 1–8, 2014.

[6] Derry FitzGerald and Jouni Paulus. Unpitched percussion transcription. In *Signal Processing Methods for Music Transcription*. Springer, 2006.

[7] Nicolai Gajhede, Oliver Beck, and Hendrik Purwins. Convolutional Neural Networks with Batch Normalization for Classifying Hi-hat, Snare, and Bass Percussion Sound Samples. In *Proc. of the Audio Mostly*, pages 111–115, 2016.

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

[9] Olivier Gillet and Gaël Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech and Language Processing*, 16(3):529–540, March 2008.

[10] Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural computation*, 18:1527–1554, 2006.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531*, pages 1–9, 2015.

[12] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167*, pages 1–11, 2015.

[13] Ping-Keng Jao and Yi-Hsuan Yang. Music Annotation and Retrieval using Unlabeled Exemplars: Correlation and Sparse Codes. *IEEE Signal Processing Letters*, 22(10):1771–1775, 2015.

[14] Diederik P. Kingma and Jimmy Lei Ba. Adam: a Method for Stochastic Optimization. In *Proc. of the International Conference on Learning Representations (ICLR)*, pages 1–15, 2015.

[15] Alexander Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. John Wiley & Sons, 2012.

[16] Jinyu Li, Rui Zhao, Jui Ting Huang, and Yifan Gong. Learning small-size DNN with output-distribution-based criteria. In *Proc. of the Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1910–1914, 2014.

[17] Jouni Paulus and Anssi Klapuri. Drum Sound Detection in Polyphonic Music with Hidden Markov Models. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009:1–9, 2009.

[18] Matthew Prockup, Erik M. Schmidt, Jeffrey Scott, and Youngmoo E. Kim. Toward Understanding Expressive Percussion Through Content Based Analysis. In *Proc. of the International Society of Music Information Retrieval Conference (ISMIR)*, 2013.

[19] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. Self-taught learning: transfer learning from unlabeled data. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 759–766, 2007.

[20] Axel Roebel, Jordi Pons, Marco Liuni, and Mathieu Lagrange. On Automatic Drum Transcription Using Non-Negative Matrix Deconvolution and Itakura Saito Divergence. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[21] Markus Schedl, Emilia Gómez, and Julián Urbano. Music Information Retrieval: Recent Developments and Applications. *Foundations and Trends® in Information Retrieval*, 8(2-3):127–261, 2014.

[22] Carl Southall, Ryan Stables, and Jason Hockman. Automatic Drum Transcription Using Bi-Directional Recurrent Neural Networks. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[23] Lucas Thompson, Matthias Mauch, and Simon Dixon. Drum Transcription via Classification of Bar-Level Rhythmic Patterns. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[24] Adam R. Tindale, Ajay Kapur, George Tzanetakis, and Ichiro Fujinaga. Retrieval of percussion gestures using timbre classification techniques. In *Proc. of the International Society of Music Information Retrieval Conference (ISMIR)*, pages 541–544, 2004.

[25] Richard. Vogl, Matthias Dorfer, and Peter Knees. Recurrent Neural Networks for Drum Transcription. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, pages 730–736, 2016.

[26] Richard Vogl, Matthias Dorfer, and Peter Knees. Drum Transcription From Polyphonic Music With Recurrent Neural Networks. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 201–205, 2017.

[27] Shinji Watanabe, Takaaki Hori, Jonathan L. Roux, and John R. Hershey. Student-Teacher Network Learning with Enhanced Features. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5275–5279, 2017.

[28] Bin Wu, Erheng Zhong, Derek Hao Hu, Andrew Horner, and Qiang Yang. SMART : Semi-Supervised Music Emotion Recognition with Social Tagging. In *SIAM Conference on Data Mining*, pages 279–287, 2013.

[29] Chih-Wei Wu and Alexander Lerch. Drum transcription using partially fixed non-negative matrix factorization with template adaptation. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, pages 257–263, 2015.

[30] Chih-Wei Wu and Alexander Lerch. On drum playing technique detection in polyphonic mixtures. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, pages 218–224, 2016.

[31] Kazuyoshi Yoshii, Masataka Goto, and Hiroshi G. Okuno. Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression. *IEEE Transactions on Audio, Speech and Language Processing*, 15(1):333–345, 2007.