AUTOMATIC DRUM TRANSCRIPTION WITH CONVOLUTIONAL NEURAL NETWORKS

C. Jacques

Analysis-Synthesis team,
STMS-UMR 9912, IRCAM, Sorbonne University, CNRS
Paris, France
celine.jacques@ircam.fr

A. Roebel

Analysis-Synthesis team,
STMS-UMR 9912, IRCAM, Sorbonne University, CNRS
Paris, France
axel.roebel@ircam.fr

ABSTRACT

Automatic drum transcription (ADT) aims to detect drum events in polyphonic music. This task is part of the more general problem of transcribing a music signal in terms of its musical score and additionally can be very interesting for extracting high level information e.g. tempo, downbeat, measure. This article has the objective to investigate the use of Convolutional Neural Networks (CNN) in the context of ADT. Two different strategies are compared. First an approach based on a CNN based detection of drum only onsets is combined with an algorithm using Non-negative Matrix Deconvolution (NMD) for drum onset transcription. Then an approach relying entirely on CNN for the detection of individual drum instruments is described. The question of which loss function is the most adapted for this task is investigated together with the question of the optimal input structure. All algorithms are evaluated using the publicly available ENST Drum database, a widely used established reference dataset, allowing easy comparison with other algorithms. The comparison shows that the purely CNN based algorithm significantly outperforms the NMD based approach, and that the results are significantly better for the snare drum, but slightly worse for both the bass drum and the hi-hat when compared to the best results published so far and ones using also a neural network model.

1. INTRODUCTION

Automatic music transcription is the task of describing a music signal in a symbolic form - a score - that contains all of the necessary information to replay the same music. Every event in a piece of music has to be characterized by musically relevant parameters like the pitch, time position, duration, and the instrument. Accordingly, the problem of music transcription can be divided into different challenges: onset detection, f0-estimation and instrument recognition. While the problem is considered as solved for monophonic signals, it is more challenging for polyphonic ones. The additivity of signals and the overlapping of partials of different notes make the task more and more complex as the number of sources increases.

A piece of music is generally performed by harmonic and percussive instruments. These instruments have different features. The spectrogram of a note is sparse in frequency, and a harmonic note has relatively few constraints with respect to its duration. On the contrary, a drum event covers a continuous part of the spectrum, but has a specific temporal response. Accordingly, different features are used to transcribe the different events. In this article we will focus on the automatic transcription of parts of the drum kit.

Automatic drum transcription is still a challenge today. Several methods have been proposed in literature and most of them can be categorised into two families: segment and classify or separate and detect. The first category segments the audio and then tries to describe what the audio segment contains. The second one separates different instruments and tries to detect onsets in the different channels.

In 2009, Paulus et al. proposed a method based on Hidden Markov Model (HMM) network in [1]. Recently, different deep learning methods have been proposed. Vogl et al. use a Recurrent Neural Network (RNN) which provides an activation function for the drum instrument (bass drum, snare drum and hi-hat) in [2]. The first study to use CNN for drum transcription has been performed in [3].

These different methods can be compared easily as most of them have been evaluated on the same database, the ENST drum database [4]. In light of the results, most DNN approaches seem to lag behind those using Hidden Markov Models (HMM) such as proposed in [1].

Automatic onset detection, which consists in locating the onsets of musical events in a piece of music, is an important initial step for efficient transcription. Onset detection is frequently used as a preprocessing step for more refined transcription, as used recently in [5] for piano transcription, and in [6] for drum transcription. A successful detection of all onsets significantly reduces the processing time of the subsequent transcription algorithm which does not need to be run over the complete signal.

There exists a large multitude of approaches that have been developed for the onset detection problem. Bello et al. provide a rather extensive overview of the various methods in [7]. The methods generally are variations of the following approach: after a pre-processing step, which highlights some properties of the signal facilitating the subsequent detection stage, the so called Onset Detection Function (ODF) is calculated. The local maxima of the ODF with a value above a threshold (which is a parameter of the algorithm) are then retained as onsets. Elowsson in [8] for example used the spectral flux, which is the difference of energy between the actual temporal frame and the previous one, to calculate the ODF. Many other approaches to calculate the ODF have been discussed in the literature.

Recently, onset detection methods based on deep learning have shown very good results. While some works aim to improve peak picking from an onset detection function as in [9], others use RNN (Recursive Neural Network) as in [10] to create the ODF. In 2014, Schlüter et al. investigated using CNN (Convolutional Neural Network) [11] for the onset detection task, and according to MIREX 2017[1] the CNN based onset detection can now be considered as state of the art. In [11] it is shown that the weights of the kernels of the convolutive layers that are used to detect percussive and

[1] http://nema.lis.illinois.edu/nema_out/mirex2017/results/aod/summary.html

harmonic onsets are rather different. This observation seems to suggest that these networks may not only be able to detect onsets, but to detect onsets for specific classes of instruments.

If we compare the CNN architecture used by Schlüter in [11] for general purpose onset detection and by Wang in [5] for piano onset detection, we find that the overall structure is very similar. However, they do not use the same data structure. Similarly how the RGB channels are accounted for in image processing, Schlüter uses as input three mel band spectrograms with the same number of bands but calculated from different STFT representations. On the contrary Wang uses just one constant Q spectrogram with a much larger number of bands.

The following paper aims to investigate the use of CNN for drum transcription. Two different approaches will be considered. First, we will use a CNN based onset detection as an initial step for subsequent drum transcription based on a recent method using non-negative matrix deconvolution [6]. Here we will introduce the new idea of a detection of qualified onsets meaning onsets fulfilling additional criteria - for example onsets belonging to percussive events or drum instruments. In developing the qualification of onsets further we will investigate a CNN based drum transcription where the CNN are trained to detect individual drum instruments. The later system has strong resemblance to the approach proposed in [3]. However, instead of training a multi label system that detects multiple instruments at the same time, we will separate the systems into individual drum instrument detectors. That allows us to investigate the optimal input representation for the different instruments. Instead of using the magnitude spectrogram data directly [3], we will use single and multi channel[2] mel band spectrogram data that has been introduced successfully for onset detection in [11]. We will compare two different cost functions. We will evaluate the final system using the ENST-Drums drummer that was left aside during training. That allows to compare our results with the various evaluations performed so far on the ENST-Drums database. We notably compare with results in [1] that to our knowledge are the best results reported so far. We also evaluate the available model of Southall [3] on the three drummers from ENST-Drums.

The article is organised as follows: Section 2 introduces the neural network and the different parameters to be compared, Section 3 shortly summarizes the NMD algorithm, Section 4 describes the experimental results, and finally Section 5 summarizes the conclusions and describes future work.

2. ONSET DETECTION AND COMPARISON OF CONFIGURATIONS

2.1. The CNN network

The model we use to compare different configurations is very similar to the one in [11, 5] and is represented in Figure 1. We summarize here the architecture of the network.

The input data contains mel frequency spectrogram data. The subsequent layers are alternating stacks of convolutional layers with ReLU activations and max-pooling layers. It finally ends with a fully connected layer of ReLU units and an output layer containing either a sigmoid unit or a linear unit. The output layer provides the ODF. The method then follows the standard approach to detect local maxima and uses a fixed threshold of 0.5 for the detection of onset in a given frame, which significantly simplifies the algorithmic design.

The feature maps at the output of these layers can be seen as a convolution between the input and a filter kernel. Usually in computer vision, the convolution is achieved with square filter. In time-frequency representation, the two dimensions represent two different quantities. As the aim of the network is to find changes over time dimension, it can be more interesting to use narrow rectangular filters frequency-wise and the max-pooling operations performed only on the frequency axis.

Following [11] we apply dropout with 50% drop out probability at the output of the first fully connected layer, to reduce overfitting during the training.

2.2. Parameter comparison

2.2.1. Loss function

The loss function used to direct the optimization of the neural network measures the divergence between the predicted value - the output of the network - and the target label. For onset detection, cross-entropy is commonly used because the task of detecting an onset in a frame has some relations to a binary classification task: frames containing an onset are marked as 1 and frames without onsets are marked as 0.

We note however, that the resemblance of the target ODF with a probability is only partially followed. As the CNN model is smooth in all parameters, the ODF function produced is smooth as well. Accordingly, a Dirac-impulse is difficult to produce, and therefore, similarly to [11], we will construct the target function by means of placing a sequence of three ones centered at the annotated onset. Broadening the target labels has the beneficial effect of increasing the pressure on the network to correctly represent the target labels, and at the same time reduces the problems of incoherent label positions. In our experiments we have seen that broadening the labels leads to reduced training times and slightly improved results. The use of a CNN as onset detector does not require the ODF to be confined to $[0, 1]$. This fact motivates us to compare two different cost functions combined with two corresponding output activation functions. On the one hand there is the binary cross entropy together with sigmoid activation function, and on the other hand the linear (ReLU) output unit with MSE loss function. We will discuss the results of the use of these two loss functions in the experimental section.

2.2.2. Input data structure

Kelz et al. in [12] compare the importance of hyper-parameters for piano transcription and they rank some hyper-parameters in respect to relative importance. The data representation is the second most important hyper-parameter. As a matter of fact, several different data representations are used as input data throughout the literature.

Schlüter et al. in [11] use three log-magnitude mel band spectrograms obtained with different time-frequency resolutions. They process the short time Fourier transform (STFT) with a hop size of 10 ms and window sizes of 23 ms, 46 ms and 93 ms. As the spectrograms must have the same size, they filter the spectrogram with an 80-band mel filter bank covering the band from 27.5 Hz to 16 kHz. We will subsequently denote this representation as multi

[2] the term channel will be used for the feature channels of a deep network in the following and has nothing to do with the channels of stereo audio signal

[3] `https://github.com/CarlSouthall/ADTLib`

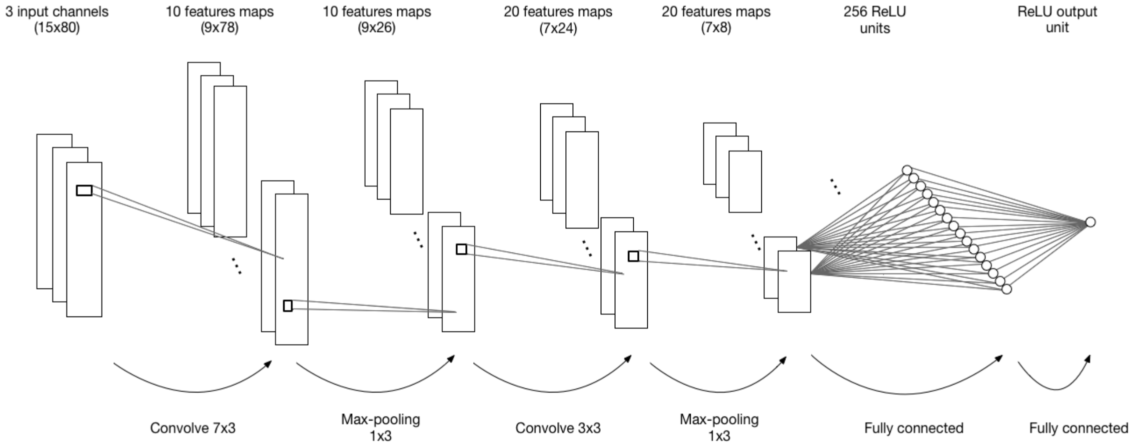

Figure 1: *Convolutional neural network used for this work.*

channel mel spectrogram (MCMS) where the term channel refers to the feature channel of a DNN.

On the contrary, Wang in [5] uses a single constant Q transform spectrogram. In this paper, we compare different data representations. We feed the eight networks with spectrograms with different resolution. Two STFT are processed with two different window sizes, 0.064 ms and 0.125 ms. Then for each spectrogram four mel spectrograms are calculated with triangular filters to compare four numbers of mel-bands: 116, 174, 231 and 289. We compare these mel spectrograms calculated from an individual STFT with the input representation proposed by Schlüter.

3. APPLICATION TO DRUM TRANSCRIPTION

We will use the CNN presented in this article in two ways to perform automatic drum transcription: combined with an ADT algorithm or alone.

As mentioned in the introduction we will investigate qualified onset detection with CNN with the objective to use these qualified onsets in the context of drum transcription. By "qualified onsets", we mean onsets that are created by one of the three targeted parts of a drum kits (hi-hat, bass drum and snare drum), either in collection (onset of any of these instruments) or individually.

In the first case, to achieve drum transcription, we combine the onset detection with a second stage to determine which of the three instruments have generated the onset. In the second case CNNs will perform the complete transcription task.

3.1. Combination of onsets detector with a drum transcription algorithm

The NMD algorithm for drum transcription we will use in the following is detailed in [6]. It decomposes the time-frequency representation of the audio signal into a convolution of a dictionary containing patterns of instruments and a matrix of activations.

For percussive instruments, the temporal response is a significant characteristic. This is the reason for using a dictionary of two-dimensional time-frequency patterns. These are previously learned from isolated events of each instrument.

The dictionary contains only patterns from drum instruments (hi-hat, snare drum and bass drum). But the drum transcription is processed on polyphonic music with harmonic instruments. To avoid the activation of drum patterns by other events, the decomposition includes some patterns in the dictionary dedicated to representing the non percussive part of the signal, which we call the background.

To reduce the computational costs, a prior knowledge of the onset position is given to the algorithm. An external algorithm, e.g. [8], feeds the transcription algorithm with the onsets that it detected. The transcription algorithm focuses on the parts of the signal that are around these positions. At these positions several instruments are likely to play. In that case, the segment study enables to separate them.

For each segment, the NMD algorithm aims to approach the studied spectrogram by activating some patterns from the dictionary. In order to, the dictionary of patterns, called W, and activations H are usually updated iteratively. For our algorithm, only background patterns in W are updated but all activations are concerned by the updating step. The update rules are calculated by minimizing a cost function, here the Itakura-Saïto divergence. As the background patterns are very flexible, they could in principle represent all parts of the signal under study. Therefore, it is important to penalize the algorithm for the use of background patterns. To this end the objective function

$$C = D_{IS}(V|\sum_t W^t H^{t\rightharpoonup}) + \lambda_{seg} P(H), \tag{1}$$

used for the decomposition contains a regularization term $P(H)$ that penalizes the use of background patterns. λ_{seg} is a weight to give more or less importance to the penalization.

To keep the non-negativity property, we use multiplicative updates:

$$W^p_{lb} \qquad \leftarrow W^p_{lb} \frac{\displaystyle\sum_n \frac{V_{ln} H_{p,n-b}}{V^2_{est_{ln}}}}{\displaystyle\sum_n \frac{H_{p,n-b}}{V_{est_{ln}}}} \tag{2}$$

$$H_{pn} \quad \leftarrow H_{pn} \frac{\sum_f \sum_t \left(W^t_{fp} \frac{V_{f(n+t)}}{(V_{est}{}_{f(n+t)})^2} \right)}{\sum_f \sum_t \left(\frac{W^t_{fp}}{V_{est}{}_{f(n+t)}} \right) + \lambda_{hseg} \mathbb{1}_{p \in bg}} \tag{3}$$

$$\tag{4}$$

with p the number of patterns, l the frequency bin and n the time frame and with bg designating the background.

Once all segments are analyzed, we follow the procedure described in [6] to adapt the detection thresholds that are applied to the activation to retain onsets of the targeted instruments.

3.2. Using CNN to transcribe drum parts

We also can use CNN described in section 2.1 to transcribe one of the targeted instrument. Instead of training the network to detect qualified onsets, we train three individual networks so that each of them detects only one instrument.

4. RESULTS

4.1. Datasets

4.1.1. RWC dataset

The training database used to adapt the CNN is extracted from the Real World Computing (RWC) music database [13]. This database contains annotated polyphonic music of different styles in MIDI format. We choose two genres, Pop and Jazz and pick only pieces where drums are present. For Jazz, there are 34 pieces of music and 100 for the Pop database. Each piece was generated with three different publicly available MIDI sound fonts: FluidR3_GM, GuGS_1.47 and HQOrchestralSFCollv2.1.2. The training database finally contains 102 jazz pieces and 300 pop pieces. In addition, we add recordings of a the single targeted instrument. These recordings are given in SMT-Drums.

For some of the experiments, evaluation is performed using a small hold out test set containing four pieces from the synthetic RWC database described above.

4.1.2. ENST-Drums dataset

The ENST-Drums database [4] is composed of different multichannel recordings from three drummers on three different drum kits. For each drummer, the data set provides individual hits and phrases, individual soli which are more complex than the phrases and longer tracks played without scores but with an accompaniment. For these longer tracks, called 'minus-one', the accompaniment is provided with two mixes: "dry" where minimal effects are added and "wet" with effects and compression. The "wet" mix sounds closer to commercial recordings than "dry" mix does and we use the "wet" mix for the following evaluation.

We use the 'minus-one' tracks mixed with the synchronized accompaniment. As in [1], scaling factors are applied to the different parts: 2/3 for the drums and 1/3 for the accompaniment. The data set also provides the ground truth annotations for each percussive instrument. The test database contains 64 tracks (21 for two drummers and 22 for the last one) which last between 30 s and 75 s.

The evaluation is performed by using the drummer cross validation procedure on the ENST-Drums database [4].

4.2. Evaluation criteria

To evaluate the algorithms, the detected onsets are compared to the ground truth onsets. A detected onset is considered correct if the absolute time difference with the associated ground truth onset does not exceed 30 ms. We denote by Tp the true positives, correctly detected onsets, by Fp false positives, detected onsets which are not in ground truth annotations and by Fn false negatives, onsets present in ground truth annotation but not detected by the algorithm.

Several measures are calculated from these values. The precision P gives the part of detected onsets which is relevant and the recall R gives the part of relevant onsets which is selected. They are defined as:

$$P = \frac{Tp}{Tp + Fp} \quad R = \frac{Tp}{Tp + Fn} \tag{5}$$

The F-measure is a compromise between recall and precision:

$$F = \frac{2PR}{P + R} \tag{6}$$

4.3. Evaluation of onset detection for drum instruments

In the first part of the evaluation we will study the performance of the CNN onset detection algorithm for detecting specific onsets. In our case, this means the onsets of any of the targeted percussive instruments. The goal of this first step is to prepare the subsequent integration of the CNN onset detection algorithm as preprocessing step into the NMD drum transcription algorithm.

Following a general practice, we will evaluate the detection of the three main instruments of the percussive part: bass drum (BD), snare drum (SD) and the hi-hat (HH). These three instruments are predominant in popular music and are representative of the rhythmic feel in music.

4.3.1. Loss function

As discussed before we compare two loss functions along with adequate changes in the output activation function: binary cross entropy with sigmoid output units and mean square error with ReLU output units. Several networks are trained with different configurations. We study four numbers of mel-bands (116, 174, 231 and 289) and two sizes of STFT window (0.064 s and 0.125 s). We also give the results for the MCMS input data configuration detailed in 2.2.2. The networks are trained and evaluated to detect the onsets of any of the three targeted percussive onsets (hi-hat, bass drum and snare drum) in the RWC database detailed in section 4.1.1.

In Figures 2 and 3, the results obtained with binary cross entropy are consistently outperforming those that are obtained with mean square error. For the following comparison, we will therefore focus on the binary cross entropy. We note that the onset prediction performance of only the three target percussive instruments is encouraging with F-measure above 90% for all configurations. There is no apparent and significant difference between any of the input data structures.

4.3.2. Evaluation the influence of data structure on the ENST-Drums database

To improve the relevance of the evaluation for real world sounds we will now evaluate CNN drum detection approach on the recordings of the ENST-Drums database [4]. The evaluation follows the common three-fold cross-validation scheme with the three configurations of the 3 drummers of the ENST-Drums database 4.1.2. The networks are trained on two drummers of the dataset and tested on the remaining one. We use all pieces available in the

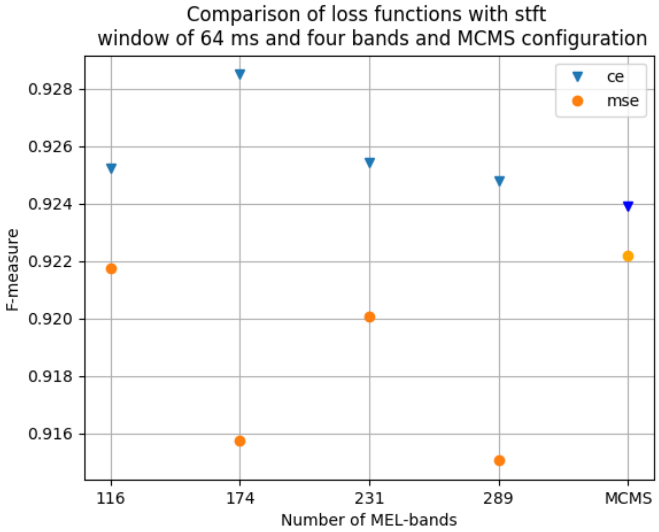

Figure 2: *Comparison of loss functions on RWC database: binary cross entropy and mean square error, for STFT window 0.064 s.*

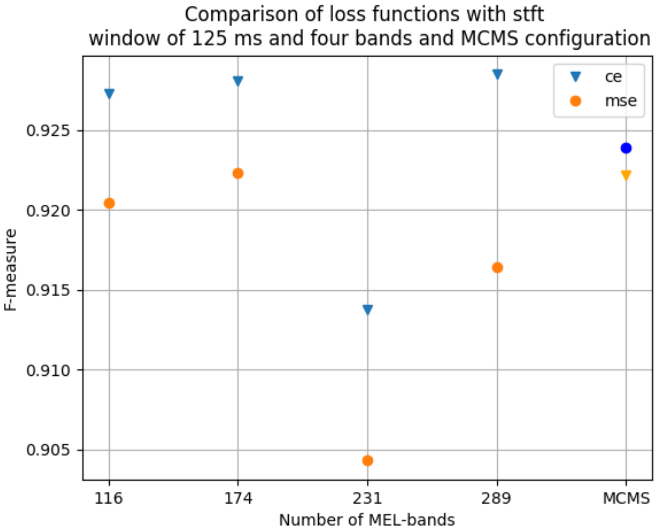

Figure 3: *Comparison of loss functions on RWC database: binary cross entropy and mean square error, for STFT window 0.125 s.*

data set for the learning phase, during which the evaluation is performed over the minus-one of the same drummers to determine the optimal result for drum detection (according to the F-measure). Then we test the generalization on the third drummer who was not used during training.

We compare the different data input configurations: two STFT window sizes 64 ms and 125 ms and four numbers of mel-bands 116, 174, 231 and 289 and the MCMS input representation. The Figure 4 averages the results over the three experiments for the detection of all percussive onset.

We notice that contrary to the evaluation with the RWC database, for the ENST-Drums database the use of the MCMS format (three spectrograms) seems to provide a significantly better results, improving the performance from 91.5% F-measure for the best single channel mel-band spectrogram to nearly 93.5% for the MCMS. While the MCMS representation was equivalent with the individual spectrogram formats on the RWC database, it is significantly better for the ENST-Drums database. That suggests the conclusion that the multiple time resolutions in the different channels of the MCMS lead to improved robustness of the final detection.

It is interesting to see to what extent the training of MCMS detector on specific onsets (the main three percussive instruments) does change its performance. To this end we use the MCMS detector provided by Schlüter in the madmom package [14]. We evaluate the two detectors on a different hold out test set of the RWC database and we find that the specific onset detector significantly improves the detection performance in F-measure from 86.8% for the general purpose onset detector to 93.2% for the percussive onset detector.

4.4. Application to drum transcription

Characterizing detected onsets might be advantageous for drum transcription. We investigate here two uses of the MCMS format for the drum transcription task. The first method combines the drum onset detector based on CNN with the ADT algorithm based on NMD described in 3.1. The CNN gives the drum onset positions and the NMD algorithm studies the segments around these positions to determine which percussive instruments provided the onset. The second one uses three individual CNNs. Each CNN is trained to detect one of the three main percussive instruments.

4.4.1. Drum onset detector combined to automatic drum transcription algorithm

Given the rather high performance of the drum onset detection algorithm, we are interested in seeing the effect of the specific onset detection when combined with an NMD based drum transcription algorithm [6]. We evaluate the performance on ENST-Drums dataset and present in Table 1 the average results on the three cross-validation experiments. We compare the obtained results with Paulus' and Southall's results. We evaluate the models by transcribing the drum parts for the 'minus one' pieces of ENST-Drums and perform the mean over the three drummers.

Table 1: *Results of transcription on three-fold cross validation.*

Methods	Metric	BD	SD	HH
HMM+ MLLR [1]	P(%)	80.2	66.3	84.7
	R(%)	81.5	45.3	82.8
	F(%)	80.8	53.9	83.6
Soft Attention+ mechanisms [15]	P(%)	98.5	88.2	67.8
	R(%)	62.2	40.1	87.9
	F(%)	72.0	53.7	76.4
NMD fed by drum onset detected by CNN	P(%)	79.6	68.8	72.6
	R(%)	64.7	43.9	67.1
	F(%)	68.9	52.6	68.3

Feeding the drum onsets to the NMD algorithm does not enable it to reach Paulus's or Southall's results.

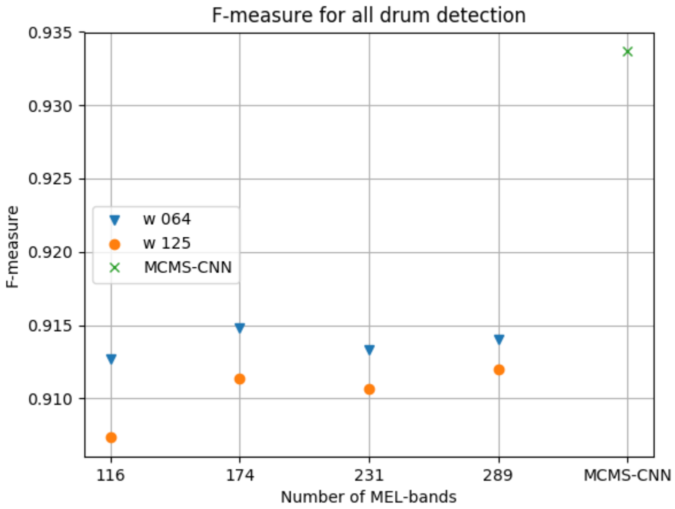

Figure 4: *Comparison different input configurations on percussive onset detection task on ENST-Drums dataset.*

4.4.2. *Individual CNNs trained on each drum instrument*

In a final experiment, motivated by the very good performance of the CNN based drum detection algorithm, we evaluate the CNN specific onset detectors trained to detect the individual drum instrument events. We therefore perform the complete transcription of an individual instrument. We focus this last experiment on the MCMS network which had the best performance in the previous experiments. Three independent CNNs are involved in this experiment. Each network is trained to detect only one of the three main percussive instrument. They are evaluated with the three-fold cross validation on the ENST-Drums database. The results averaged over the three folds of the cross validation are given in Table 2. The results of drum onset detection in 'all drums' are also displayed. They are obtained with the CNN trained to detect qualified onsets (without distinction between instrument) for our method. Southall's model does not provide those results.

Table 2: *Results of drum transcription per instrument on three-fold cross validation.*

Methods	Metric	BD	SD	HH	Percus.
HMM+	P(%)	80.2	66.3	84.7	79.0
MLLR [1]	R(%)	81.5	45.3	82.8	70.9
	F(%)	80.8	53.9	83.6	74.7
Soft Attention+	P(%)	98.5	88.2	67.8	-
mechanisms [15]	R(%)	62.2	40.1	87.9	-
	F(%)	72.0	53.7	76.4	-
CNN with	P(%)	77.5	57.9	71.0	93.7
MCMS	R(%)	75.0	67.0	89.7	93.0
configuration	F(%)	76.2	62.1	79.3	93.4

We notice that the CNN provides comparatively good results for the snare drum, for which it obtains 8pts more in F-measure than Paulus' method. But it also loses 4pts for the two other instruments, the bass drum and hi-hat. Our model is better than Southall's method.

Comparing the bass-drum results between the HMM and CNN methods we can find an explanation for the reduced performance in Table 3, which displays the results of the individual folds of the cross evaluation experiment. While the detection of drummer 3 and drummer 2 are performing very satisfyingly, the recall of drummer 1 is particularly low. Listening to the bass drum signals of the different drummers reveals that the bass drum signal of drummer 1 is clearly different from the two other drummers. Its energy is significantly lower in comparison to the bass drum signals of drummers 2 and 3. We have tried to counter this difference by means of using different mixes when training the network, without achieving any improvement. One can also observe that the bass drum signal of drummer 1 contains a much less pronounced onset, which may constitute another explanation for the low recall. Here, the high specificity of the CNN leads to an over-fitting of the training signals, which in turn reduces the recall for drummer 1. Although Southall's model seems to encounter the same problem, the HMM model displayed in Table 2 apparently does not have the same issue with drummer 1. It may indicate that the CNN model we chose and which worked very well for the general drum detection task, is too complex.

Table 3: *Results of bass drum transcription on three-fold cross validation.*

Train drummers	Eval drummer	P	R	F
1 and 2	3	82.5	96.7	89.1
2 and 3	1	75.1	36.7	45.0
3 and 1	2	74.6	98.1	84.8

An other idea to improve detection of bass drum played by drummer 1 is to normalize over time only. It highlights the sudden changes of energy which can be characteristic of onsets. However, this kind of normalization modify the relation of energy between the frequency bands. But as the energy of bass drum is located in low frequency bands, the networks is able to correctly detect the onsets. In fact, for drummer 1, the F-measure on drummer 1 for bass drum raises 67.7 % instead of 45%. The results for the other drum instruments and for the percussive instruments are given in Table 4. We compare the results with [1] and [15].

Table 4: *Results of drum transcription per instrument on three-fold cross validation with normalization over time.*

Methods	Metric	BD	SD	HH	Percus.
HMM+	P(%)	80.2	66.3	84.7	79.0
MLLR [1]	R(%)	81.5	45.3	82.8	70.9
	F(%)	80.8	53.9	83.6	74.7
Soft Attention+	P(%)	98.5	88.2	67.8	-
mechanisms [15]	R(%)	62.2	40.1	87.9	-
	F(%)	72.0	53.7	76.4	-
CNN with	P(%)	84.0	54.2	71.9	93.8
MCMS config.	R(%)	80.7	68.1	86.6	91.7
and tnorm	F(%)	81.5	59.4	77.8	92.7

The F-measure is slightly better for bass drum and significantly better for snare drum than F-measures obtained with the method of HMM. The detection of percussive onsets is also largely more effective. Although normalization over time degrades a little bit the F-measure for snare drum and hi-hat detection in comparison with our model, it is much better for bass drum detection.

5. CONCLUSIONS

In this paper, we investigated different new approaches to the use of Convolutional Neural Networks for automatic drum transcription. We compared different loss functions and input representations. We found that the best results are obtained with the MCMS representation of the input data, namely three log-magnitude spectrograms with three different STFT window sizes: 23, 46 and 93 ms filtered into 80 mel frequency bands. We trained the network for the detection of percussive onsets, achieving very good detection performance well above 90% in F-measure. The combination of specific onset detectors based on CNN with a drum (bass drum, snare drum and hi-hat) transcription algorithm based on Non-negative Matrix Deconvolution did not lead to competitive performances.

Finally, we trained three individual CNNs: each of them detecting one of the three percussive instruments (bass drum, snare drum and hi-hat). The results obtained are significantly better than the results obtained with the NMD, which leads us to believe that the use of CNN for drum transcription has more potential than the use of a non-negative decomposition. We conjecture that the main reason for the better results is the fact that the CNN is trained with an objective function (the ODF) that is much closer to the final task than the objective function used in the NMD training. Further investigation is required to compare the single label detector proposed in the present paper with the multi label detector. While the single label detector may have the advantage of specializing more on the specific instrument, it also may be the reason for the over-fitting observed notably during the bass drum detection of drummer 1 of the ENST-Drums database.

6. REFERENCES

[1] Jouni Paulus and Anssi Klapuri, "Drum sound detection in polyphonic music with hidden markov models," *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[2] Richard Vogl, Matthias Dorfer, and Peter Knees, "Drum transcription from polyphonic music with recurrent ,eural networks," *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[3] Carl Southall, Ryan Stables, and Hockman Jason, "Automatic drum transcription for polyphonic recordings using soft attention mechanisms and convolutional neural networks," *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[4] Olivier Gillet and Gaël Richard, "Enst-drums: an extensive audio-visual database for drum signals processing," *Proceedings of the 7th International Society for Music Information Retrieval Conference (ISMIR)*, 2006.

[5] Qi Wang, Ruohua Zhou, and Yonghong Yan, "A two stage approach to note-level transcription of a specific piano," *Applied Science*, 2017.

[6] Axel Roebel, Jordi Pons, Marco Liuni, and Mathieu Lagrange, "On automatic drum transcription using non-negative matrix deconvolution and itakura-saito divergence," *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 414–418, 2015.

[7] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler, "A tutorial on onset detection in musical signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[8] Anders Elowsson and Anders Friberg, "Modelling perception of speed in music audio," *Proceedings of the Sound and Music Computing Conference*, 2013.

[9] Matija Marolt, Alenka Kavcic, and Marko Provosnik, "Neural networkd for note onset detection in piano music," *Proceedings of the International Computer Music Conference (ICMC)*, 2002.

[10] Sebastian Böck, Andreas Artz, Florian Krebs, and Markus Shedl, "Online real-time onset detection with recurrent neural networks," *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx)*, September 2012.

[11] Jan Schlüter and Sebastian Böck, "Improved musical onset detection with convolutional neural networks," *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[12] Rainer Kelz, Matthias Dorfer, Filip Korzeniowski, Sebastian Böck, Andreas Artz, and Ghehard Widmer, "On the potential of simple framewise approaches to piano transcription," *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[13] Masataka Goto, Hiroki Hashigichi, Takuichi Nishimura, and Ryuichi Oka, "RWC music database: Popular, classical and jazz music databases.," *Proceedings of the 3rd International Society on Music Information Retrieval Conference (ISMIR)*, vol. 2, pp. 287–288, 2002.

[14] Sebastian Böck, Filip Korzeniowski, Jan Schüter, Florian Krebs, and Gerhard Widmer, "Madmom: a new python audio and music signal processing library," *Late-Breaking Demo Session of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[15] Carl Southall, Nicholas Jillings, Ryan Stables, and Jason Hockman, "Adtweb: An open source browser based automatic drum transcription system," *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.