Enhanced peak picking for onset detection with recurrent neural networks

Sebastian Böck[1], Jan Schlüter[2], and Gerhard Widmer[12]

[1] Dept. of Computational Perception, Johannes Kepler University, Linz, Austria
[2] Austrian Research Institute for Artificial Intelligence, Vienna

Abstract. We present a new neural network based peak-picking algorithm for common onset detection functions. Compared to existing hand-crafted methods it yields a better performance and leads to a much lower number of false negative detections. The performance is evaluated on basis of a huge dataset with over 25k annotated onsets and shows a significant improvement over existing methods in cases of signals with previously unknown levels.

1 Introduction

Onset detection is the process of finding the starting points of all musically relevant events in an audio stream. Having reliable algorithms is crucial for a lot of higher level tasks for music information retrieval, such as beat-tracking, score following, and music transcription.

Many different methods have been proposed and evaluated over the years. Comprehensive overviews of onset detection methods were presented by Bello et al. [1], Collins [5] and Dixon [6]. The spectral flux method turned out to be a good overall performer, and is used as a basis for this work. The final onsets in all these detection functions are obtained with basic peak selections algorithms, which are mostly hand-crafted methods based on psychoacoustic theory and heuristics. Rosao et al. [8] investigated the influence of different peak-picking methods on the performance of onset detection methods and found that the right choice of the parameters depends on the type of onsets present in the audio.

We address this issue by applying a new peak-picking method to established onset detection functions, which learns its parameters on the basis of a huge annotated dataset that covers a huge variety of sounds, music and onset types.

2 Onset detection function

As a basis for our experiments, we chose the computationally inexpensive and well performing *SuperFlux* onset detection function [4], an enhanced version of the spectral flux algorithm. It achieves a level of performance superior to any other non-probabilistic approach and close to the current state-of-the-art onset detection method called *OnsetDetector* [7].

First the audio signal is transferred from the time domain to the frequency domain with the Short-Time Fourier Transform (STFT). Each frame of the

In MML 2013: International Workshop on Machine Learning and Music, ECML/PKDD, Prague, 2013.

magnitude spectrogram is then filtered with a filterbank with 80 overlapping triangular filters which are spaced equally on a logarithmic frequency scale. This pre-processing is very close to the human ear and was found to outperform various other techniques [3]. The reduction function calculates the difference of a frame to its maximum-filtered predecessor, and sums up all positive changes to form the onset detection function.

The original implementation uses a peak detection function, which is based on calculating a moving maximum and average. The final onsets are selected, if the current value of the onset detection function is equal to the moving maximum and exceeds the moving average by at least a certain amount – the threshold parameter.

3 Neural network based peak detection

Instead of this hand-crafted approach, we propose a more universal approach based on a recurrent neural network (RNN), which is trained in a supervised manner. The recurrent connections in the fully connected hidden layers enable the neural network to model the temporal context of the onset detection function. The used network has a single linear input neuron and two hidden layers with four *tanh* units each. For the output layer a single neuron is used as a binary classifier with a constant decision threshold of 0.25 (this value was determined as the mean value yielding the best F-measure on the training set over all 8 training folds) and all values exceeding this threshold are considered as onsets.

We propose two different neural network topologies. When used in online scenarios (i.e., the strictly causal processing of a continuous audio stream) where no future information is available, the topology described above with 65 weighted connections is used. For offline mode (where future information can be exploited), the hidden layers of the network are doubled in a bidirectional way to form a bidirectional recurrent neural network (BRNN) [9]. This way the network can model the temporal context of the onset in both directions. The topology is shown in Figure 1. During training, the inputs and targets are presented to the second set of hidden layers in reverse temporal order. This topology has a total of 161 weights.

All weights of the networks are initialized randomly with a Gaussian distribution having 0 mean and 0.1 standard deviation. The network is trained with gradient descent and backpropagation of the errors through time. For training, the complete dataset is split into eight disjoint sets. Six are used for training, one used a validation set to perform early stopping to prevent overfitting to the training data. The remaining set is used for testing.

4 Evaluation

For evaluation, we use the dataset used in [3]. The 321 audio excerpts consists mostly of mixed audio material covering different types of musical genres, performed on various instruments. They have a total length of approximately 102 minutes and 25,927 annotated onsets. All results given in this section are obtained with 8-fold cross-validation.

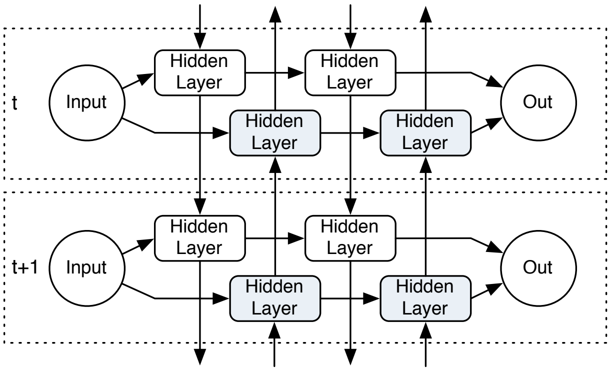

Fig. 1. Bidirectional recurrent neural network (BRNN) topology used for peak picking. Each fully connected recurrent hidden layer has 4 *tanh* units. The second set of bidirectional hidden layers (shaded) are only connected in offline mode.

Performance is evaluated regarding Precision, Recall, and the F-measure. A reported onset is considered as detected correctly if there is a ground truth annotation within the evaluation window of 50 ms (±25 ms) around the detected position. Any additionally reported onsets are counted as false positives and additional targets are treated as false negatives respectively.

We compare the results with the original *SuperFlux* implementation and the current state-of-the-art algorithms for online [2] and offline [7] onset detection. Table 1 shows that the new peak detection algorithm is able to outperform the original hand-crafted detection method both in online and offline mode. Especially the higher recall rates indicate that the new peak picking algorithm is able to better adapt the thresholds dynamically and thus captures more onsets in the same detection function. Interestingly, in online mode the new method achieves even better performance that the current state-of-the-art implementation, although it uses a much simpler neural network.

The positive effect of the new peak-picking method on the performance can be better seen if the signal is attenuated by 15 dB. This test was also performed in [3], and simulates how good an algorithm can deal with varying signal levels of unknown magnitude (e.g., a live microphone input signal). Although the original *SuperFlux* performs already much better than all algorithms in [3], the new peak-picking method shows a significant performance boost for this scenario. Only the computationally more expensive purely neural network based approach [2] does not show any performance impact in this case.

5 Conclusion

In this paper we presented a new peak-picking method for existing onset detection functions, which is able to outperform existing hand-crafted methods and which shows a significant performance boost in cases of unknown signal levels.

	Precision	Recall	F-measure
offline			
OnsetDetector.2012 [7]	0.892	0.855	0.873
SuperFlux [4]	0.883	0.793	0.836
SuperFlux w/ new peak detection	0.890	0.822	0.854
online			
OnsetDetectorLL [2]	0.863	0.783	0.821
SuperFlux [4]	0.855	0.787	0.820
SuperFlux w/ new peak detection	0.857	0.804	0.830
online (signal -15 dB)			
OnsetDetectorLL [2]	0.862	0.780	0.819
SuperFlux [4]	0.952	0.616	0.748
SuperFlux w/ new peak detection	0.933	0.703	0.802

Table 1. Precision, Recall and F-measure of different onset detection algorithms using offline and online peak detection (with additional signal attenuation).

6 Acknowledgments

This work is supported by the European Union Seventh Framework Programme FP7 / 2007-2013 through the PHENICX project (grant agreement no. 601166) and the Austrian Science Fund (FWF) through project TRP 307-N23.

References

1. J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5), 2005.
2. S. Böck, A. Arzt, F. Krebs, and M. Schedl. Online real-time onset detection with recurrent neural networks. In *Proc. of the 15th International Conference on Digital Audio Effects*, 2012.
3. S. Böck, F. Krebs, and M. Schedl. Evaluating the online capabilities of onset detection methods. In *Proc. of the 13th International Society for Music Information Retrieval Conference*, 2012.
4. S. Böck and G. Widmer. Maximum filter vibrato suppression for onset detection. In *Proc. of the 16th International Conference on Digital Audio Effects*, 2013.
5. N. Collins. A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. In *Proc. of the AES Convention 118*, 2005.
6. S. Dixon. Onset detection revisited. In *Proc. of the 9th International Conference on Digital Audio Effects*, 2006.
7. F. Eyben, S. Böck, B. Schuller, and A. Graves. Universal onset detection with bidirectional long short-term memory neural networks. In *Proc. of the 11th International Society for Music Information Retrieval Conference*, 2010.
8. C. Rosão, R. Ribeiro, and D. Martins De Matos. Influence of peak selection methods on onset detection. In *Proc. of the 13th International Society for Music Information Retrieval Conference*, 2012.
9. M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 1997.