MIREX 2010 SUBMISSION: ONSET DETECTION WITH BIDIRECTIONAL LONG SHORT-TERM MEMORY NEURAL NETWORKS

Sebastian Böck, Florian Eyben, Björn Schuller
Institute for Human-Machine Communication
Technische Universität München
sb@minimoog.org, eyben@tum.de, schuller@tum.de

ABSTRACT

We present a new onset detector applicable for all kinds of music, including complex music mixes. It is based on auditory spectral features and relative spectral differences processed by a bidirectional Long Short-Term Memory recurrent neural network, which acts as reduction function. The network is trained with a large database of onset data covering various genres and onset types. Due to the data driven nature, our approach does not require the onset detection method and its parameters to be tuned to a particular type of music.

1. ALGORITHM DESCRIPTION

This section describes our algorithm for onset detection in music signals, which is based on bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks [3, 4]. The approach is purely data driven and is able to model the context an onset occurs in. The properties of an onset and the amount of relevant context are thereby learned from the data set used for training. The audio data is transformed to the frequency domain via two parallel STFTs with different window sizes. The obtained magnitude spectra and their first order differences are used as inputs to the BLSTM network, which produces an onset activation function at its output. Figure 1 shows this basic signal flow. The individual blocks are described in more detail in the following sections.

1.1 Feature extraction

As input, the raw PCM audio signal with a sampling rate of $f_s = 44.1\,\text{kHz}$ is used. To reduce the computational complexity, stereo signals are converted to a monaural signal by averaging both channels. The discrete input audio signal $x(t)$ is segmented into overlapping frames of W samples length ($W = 1024$ and $W = 2048$), which are sampled at a rate of one per 10 ms (onset annotations are available on a frame level). A Hamming window is applied to these frames. Applying the STFT yields the complex spectrogram $X(n, k)$, with n being the frame index, and k the frequency bin index. The complex spectrogram is converted to the power spectrogram $S(n, k) = |X(n, k)|^2$.

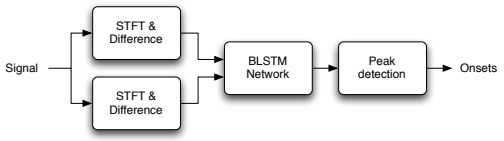

Figure 1. Basic signal flow of the new neural network based onset detector

The dimensionality of the spectra is reduced by applying psychoacoustic knowledge: a filterbank with 40 triangular filters, which are equidistant on the Mel scale, is used to transform the spectrogram $S(n, k)$ to the Mel spectrogram $M(n, m)$. To match human perception of loudness, a logarithmic representation is chosen:

$$M^{log}(n, m) = log\left(M(n, m) + 1.0\right) \quad (1)$$

The positive first order difference $D^+(n, m)$ is calculated by applying a half-wave rectifier function $H(x) = \frac{x+|x|}{2}$ to the difference of two consecutive Mel spectra:

$$D^+(n, m) = H\left(M^{log}(n, m) - M^{log}(n-1, m)\right) \quad (2)$$

1.2 Neural Network stage

As a neural network, an RNN with BLSTM units is used. As inputs to the neural network, two log Mel-spectrograms $M_{23}^{log}(n, m)$ and $M_{46}^{log}(n, m)$ (computed with window sizes of 23.2 ms and 46.4 ms) and their corresponding positive first order differences $D_{23}^{+}(n, m)$ and $D_{46}^{+}(n, m)$ are applied, resulting in 160 input units. The network has three hidden layers for each direction (6 layers in total) with 20 LSTM units each. The output layer has two units, whose outputs are normalised to both lie between 0 and 1, and to sum to 1, using the softmax function. The normalised outputs represent the probabilities for the classes 'onset' and 'no onset'. This allows the use of the cross entropy error criterion to train the network [3].

1.2.1 Network training

For network training, supervised learning with early stopping is used. The used training set consists of 110 audio excerpts with lengths between 1.4 and 60 seconds taken from the data set introduced by Bello in [1] and from the ISMIR 2004 tempo induction contest [1] [2]. A part of the annotation work was done by Lacoste and Eck for their neural network approach [2]. The remaining parts were manually labelled by an expert musician [3]. All annotations have been revised for network training.

Each audio sequence is presented frame by frame (in correct temporal order) to the network. Standard gradient descent with backpropagation of the output errors is used to iteratively update the network weights. To prevent overfitting, the performance (cross entropy error, cf. [3]) on a separate validation set is evaluated after each training iteration (epoch). If no improvement of this performance over 20 epochs is observed, the training is stopped and the network with the best performance on the validation set is used as the final network. The gradient descent algorithm requires the network weights to be initialised with non zero values. We initialise the weights with a random Gaussian distribution with mean 0 and standard deviation 0.1.

1.3 Peak detection stage

A network obtained after training as described in the previous section is able to classify each frame into two classes: 'onset' and 'no onset'. The standard method of choosing the output node with the highest activation to determine the frame class has not proven effective. Hence, only the output activation of the 'onset' class is used. Thresholding and peak detection is applied to it, which is described in the following sections:

1.3.1 Thresholding

One problem with existing magnitude based reduction functions is that the amplitude of the detection function depends on the amplitude of the signal or the magnitude of its short time spectrum. Thus, to successfully deal with high dynamic ranges, adaptive thresholds must be used when thresholding the detection function prior to peak picking. Similar to phase based reduction functions, the output activation function of the BLSTM network is not affected by input amplitude variations, since its value represents a probability of observing an onset rather than representing onset strength. In order to obtain optimal classification for each song, a fixed threshold θ is computed per song proportional to the median of the activation function (frames $n = 1 \ldots N$), constrained to the range from $\theta_{min} = 0.1$ to $\theta_{max} = 0.3$:

$$\theta^* = \lambda \cdot \text{median}\{a_o(1), \ldots, a_o(N)\} \quad (3)$$

$$\theta = \min\left(\max\left(0.1, \theta^*\right), 0.3\right) \quad (4)$$

[1] http://mtg.upf.edu/ismir2004/contest/tempoContest/node5.html

[2] http://w3.ift.ulaval.ca/~allac88/dataset.tar.gz

[3] Data available at: http://mir.minimoog.org/

with $a_o(n)$ being the output activation function of the BLSTM neural network for the onset class, and the scaling factor λ chosen to maximise the F_1-measure on the validation set. The final onset function $o_o(n)$ contains only the activation values greater than this threshold:

$$o_o(n) = \begin{cases} a_o(n) & \text{for } a_o(n) > \theta \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

1.3.2 Peak picking

The onsets are represented by the local maxima of the onset detection function $o_o(n)$. Thus, using a standard peak search, the final onset function $o(n)$ is given by:

$$o(n) = \begin{cases} 1 & \text{for } o_o(n-1) \leq o_o(n) \geq o_o(n+1) \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

2. RESULTS

The described OnsetDetector performed best in the MIREX 2010 evaluation (see Table 1). This was achieved without any optimisation on the onset types in test data.

Onsets [%]	F-Measure	Precision	Recall
BES1	78.66	76.25	84.69
AR3	77.66	84.67	75.15
TZC1	74.37	75.58	76.97
BT2	73.36	77.95	71.63
ME1	67.09	75.81	63.78
TGGL1	59.42	65.27	59.82
RVCCR1	43.84	70.24	37.37

Table 1. Results for the MIREX 2010 onset detection evaluation. Only the best performing algorithm of other participants/groups are shown.

3. 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):1035–1047, Sept. 2005.

[2] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1832–1844, Sept. 2006.

[3] A. Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. PhD thesis, Technische Universität München. Munich, Germany. 2008.

[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computing*, 9(8):1735–1780, 1997.