UNIVERSAL ONSET DETECTION WITH BIDIRECTIONAL LONG SHORT-TERM MEMORY NEURAL NETWORKS

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

Alex Graves
Institute for Computer Science VI
Technische Universität München
graves@in.tum.de

ABSTRACT

Many different onset detection methods have been proposed in recent years. However those that perform well tend to be highly specialised for certain types of music, while those that are more widely applicable give only moderate performance. In this paper we present a new onset detector with superior performance and temporal precision 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. We compare results on the Bello onset data set and can conclude that our approach is on par with related results on the same set and outperforms them in most cases in terms of F_1-measure. For complex music with mixed onset types, an absolute improvement of 3.6% is reported.

1. INTRODUCTION

Finding onset locations is a key part of segmenting and transcribing music, and therefore forms the basis for many high level automatic retrieval tasks. An onset marks the beginning of an acoustic event. In contrast to music information retrieval studies which focus on beat and tempo detection via the analysis of periodicities (e. g. [7, 9]), an onset detector faces the challenge of detecting single events, which need not follow a periodic pattern. Recent onset detection methods (e. g. [5, 16, 17]) have matured to a level where reasonable robustness is obtained for polyphonic music. However, the methods are specialised or tuned to specific kinds of onsets (e. g. pitched or percussive) and lack the ability to perform well for music with mixed onset types. Thus, multiple methods need to be combined or a method has to be selected depending on the type of onsets to be analysed.

In this paper we propose a novel, robust approach to onset detection, which can be applied to any type of music. Our approach is based on auditory spectral features and Long Short-Term Memory (LSTM) [13] recurrent neural networks. The approach is purely data driven, and as we will see, yields a very high temporal precision as well as detection accuracy.

The rest of this paper is structured as follows. A brief overview of the state of the art in onset detection is given in Section 2, and Section 3 provides an introduction to LSTM neural networks. Section 5 describes the Bello onset data set [2] as well as introducing a new data set. Experimental results for both data sets are provided in Section 6, along with a comparison to related systems.

2. EXISTING METHODS

Most onset detection algorithms are based on the three step model shown in Figure 1. Some methods include a preprocessing step. The aim of preprocessing is to emphasise relevant parts of the signal. Next, a reduction function is applied, to obtain the detection function. This is the core component of an onset detector. Some of the most common reduction functions found in the literature are summarised later in this section.

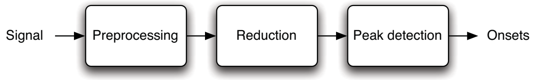

Figure 1. Traditional onset detection workflow

The last stage is to extract the onsets from the detection function. This step can be subdivided into post processing (e. g. smoothing and normalising of the detection function), thresholding, and peak picking. If fixed thresholds are used, the methods tend to pick either too many onsets in louder parts, or miss onsets in quieter parts. Hence, adaptive thresholds are often used. Finally the local maxima above the threshold, which correspond to the detected onsets, are identified by a peak picking algorithm.

Early reduction functions, such as [14], operated in the time domain. This approach normalises the loudness of the signal before splitting it into multiple bands via bandpass filters. Onsets are then detected in each band as peaks in the first order difference of the logarithm of the amplitude envelope. These band-wise onsets are then combined to yield the final set of detected onsets. More recent systems employ spectral domain reduction functions. We describe the most common ones in the following paragraphs.

2.1 Spectral domain reduction functions

Since onsets are often masked in the time domain by higher energy signals, many reduction functions operate on a spectral representation of the audio signal. The methods listed below are all based on a short-time Fourier transform (STFT) of the signal.

2.1.1 High Frequency Content

Percussive sounds have a high energy in the upper frequency bands. This is exploited by weighting each STFT bin with a factor proportional to its frequency. Summing all weighted bins yields a measure called the high frequency content (HFC), which is used as a detection function. Although this method works well for percussive onsets, it shows weaknesses for other onset types [2].

2.1.2 Spectral difference

For computation of the spectral difference function (SD), the difference of two consecutive short-time spectra is computed bin by bin. All positive differences are then summed up across all bins. Some approaches use the L_2-norm [2] for calculating the difference, whereas others use the L_1-norm [5], in which case the function is referred to as spectral flux (SF). Onset detection methods based on these methods are among the best overall performers so far.

2.1.3 Phase deviation

The methods mentioned so far rely on the spectral magnitudes. In [2] a method utilising phase information is described. The change of the phase in a STFT frequency bin is a rough estimate of its instantaneous frequency. A change of this frequency is an indicator of a possible onset. To reduce the chance of a missed onset due to phase wrap around, the mean phase change over all frequency bins is used. Dixon proposes an improvement to the phase deviation (PD) detection function called normalised weighted phase deviation (NWPD) [5], where each frequency bin's contribution to the phase deviation function is weighted by its magnitude. The result is normalised by the sum of the magnitudes.

2.1.4 Complex Domain

Another way to incorporate both magnitude and the phase information is proposed in [6]. First, the expected amplitude and phase is calculated for the current frame based on the two previous frames, assuming constant amplitude and phase change rate. The sum of the magnitude of the complex differences between the actual values for each frequency bin and the estimated values is then computed and used as a detection function. A variant of this method is called the rectified complex domain (RCD) [5]. Observing that increases of the signal amplitude are generally more relevant than decreases for onset detection, RCD modifies the original algorithm by only summing over positive amplitude changes.

2.2 Probabilistic reduction functions

An alternative approach is to base the description of signals on probabilistic models. The negative log-likelihood method [1] defines two different statistical models and observes whether the signal follows the first or the second model. A sudden change from the first model to the second can be an indication of an onset. This method shows good results for music with soft onsets, e. g. non-percussive sounds [2].

2.3 Pitch-based onset detection techniques

Collins describes an onset detection function based on a pitch detector front-end [4]. Zhou presented a combination of pitch and energy based detection functions [17]. In principle pitch-based onset detection is based on identification of discontinuities and perturbations in the pitch contour, which are assumed to be indicators of onsets.

2.4 Data-driven reduction functions

To build general detection functions, which are capable of detecting onsets in a wider range of audio signals, classifier based methods emerged. In [15] an onset detection algorithm based on a feed forward neural network, namely a convolutional neural network, is described. This system performed best in the MIREX 2005 audio onset detection evaluation.

3. NEURAL NETWORKS

Motivated by the high performance of the onset detection method of Lacoste and Eck, we investigate a novel artificial neural network (ANN) based approach. Instead of a simple feed forward neural network we use a bidirectional recurrent neural network with Long Short-Term Memory [13] hidden units. Such networks were proven to work well on other audio detection tasks, such as speech recognition [10].

This section gives a short introduction to ANN with a focus on bidirectional Long Short-Term Memory (BLSTM) networks, which are used for the proposed onset detector.

3.1 Feed forward neural networks

The most commonly used form of feed forward neural networks (FNN) is the multilayer perceptron (MLP). It consists of a minimum of three layers, one input layer, one or more hidden layers, and an output layer. All connections feed forward from one layer to the next without any backward connections. MLPs classify all input frames independently. If the context a frame is presented in is relevant, this context must be explicitly fed to the network, e. g. by using a fixed width sliding window, as in [15].

3.2 Recurrent neural networks

Another technique for introducing past context to neural networks is to add backward (cyclic) connections to FNNs. The resulting network is called a recurrent neural network (RNN). RNNs can theoretically map from the entire history of previous inputs to each output. The recurrent connections form a kind of memory, which allows input values to persist in the hidden layer(s) and influence the network output in the future. If future context is also necessary required, a delay between the input values and the output targets can be introduced.

3.3 Bidirectional recurrent neural networks

A more elegant incorporation future context is provided by bidirectional recurrent networks (BRNNs). Two separate hidden layers are used instead of one, both connected to the same input and output layers. The first processes the input sequence forwards and the second backwards. The network therefore has always access to the complete past and the future context in a symmetrical way, without bloating the input layer size or displacing the input values from the corresponding output targets. The disadvantage of BRNNs is that they must have the complete input sequence at hand before it can be processed.

3.4 Long Short-Term Memory

Although BRNNs have access to both past and future information, the range of context is limited to a few frames due to the vanishing gradient problem [11]. The influence of an input value decays or blows up exponentially over time, as it cycles through the network with its recurrent connections and gets dominated by new input values.

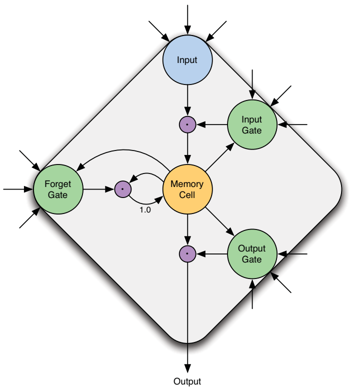

Figure 2. An LSTM block with one memory cell

To overcome this deficiency, a method called Long Short-Term Memory (LSTM) was introduced in [13]. In an LSTM hidden layer, the nonlinear units are replaced by LSTM memory blocks (Figure 2). Each block contains one or more self connected linear memory cells and three multiplicative gates. The internal state of the cell is maintained with a recurrent connection of constant weight 1.0. This connection enables the cell to store information over long periods of time. The content of the memory cell is controlled by the multiplicative input, output, and forget gates, which – in computer memory terminology – correspond to write, read, and reset operations. More details on the training algorithm employed, and the bidirectional LSTM architecture in general can be found in [10].

4. PROPOSED APPROACH

This section describes our novel approach for onset detection in music signals, which is based on bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks. In contrast to previous approaches it 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 3 shows this basic signal flow. The individual blocks are described in more detail in the following sections.

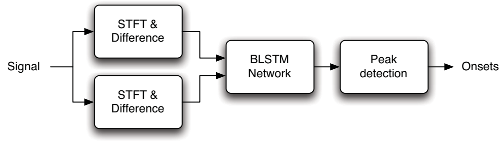

Figure 3. Basic signal flow of the new neural network based onset detector

4.1 Feature extraction

As input, the raw PCM audio signal with a sampling rate of $f_s = 44.1$ 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 $W = 2048$, see Section 4.2), 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 $S(n, k) = |X(n, k)|^2$.

The dimensionality of the spectra is reduced by applying psychoacoustic knowledge: a conversion to the Mel-frequency scale is performed with openSMILE [8]. 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:

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

4.2 Neural Network stage

As a neural network, an RNN with BLSTM units is used. As inputs to the neural network, two log Mel-spectrograms $M_{log}^{23}(n,m)$ and $M_{log}^{46}(n,m)$ (computed with window sizes of 23.2 ms and 46.4 ms ($W = 1024$ and $W = 2048$ samples), respectively) and their corresponding positive first order differences $D_{23s}^{+}(n,m)$ and $D_{46s}^{+}(n,m)$ 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 [10]. Alternative networks with a single output, where a value of 1 represents an onset frame and a value of 0 a non-onset frame, which are trained using the mean squared output error as criterion, were not as successful.

4.2.1 Network training

For network training, supervised learning with early stopping is used. 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 over-fitting, the performance (cross entropy error, cf. [10]) 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. The training data, as well as validation and test sets are described in Section 5.

4.3 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:

4.3.1 Thresholding

One problem with existing magnitude based reduction functions (cf. Section 2) 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 $\theta_{max} = 0.3$:

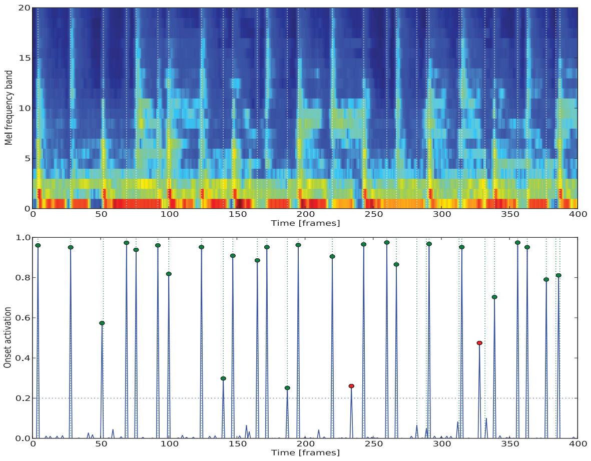

Figure 4. *Top*: log Mel-spectrogram with ground truth onsets (vertical dashed lines). *Bottom*: network output with detected onsets (marked by dots), ground truth onsets (dotted vertical lines), and threshold θ (horizontal dashed line). 4 s excerpt from 'Basement Jaxx - Rendez-Vu'.

$$\begin{aligned} \theta^{*} &= \lambda \cdot \text{median}\{a_o(1), \ldots, a_o(N)\} && (3) \\ \theta &= \min\left(\max\left(0.1, \theta^{*}\right), 0.3\right) && (4) \end{aligned}$$

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:

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

4.3.2 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:

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

5. DATA SETS

We evaluate our onset detector using the data set introduced by Bello in [2], which consists of 23 sound excerpts with lengths ranging from a few seconds to one minute (cf. Table 1). the data set is divided into four categories: pitched percussive (PP), pitched non-percussive (PNP), non-pitched percussive (NPP), and complex music mixes (MIX). The set includes audio synthesised from MIDI files as well as original recordings.

In order to effectively train the BLSTM network, the onset annotations had to be corrected in a few places: missing onsets were added and onsets in polyphonic pieces were properly aligned to match the annotation precision of the MIDI based samples. For rule-based onset detection approaches, minor inaccuracies of a few frames are not crucial since these are levelled out by the detection window during evaluation. For the BLSTM network, however, it is necessary to have temporally precise data for training. Nonetheless, the original, unmodified transcriptions are used for evaluation, to ensure a fair comparison.

To increase the size of the training data set, 87 10 s excerpts of ballroom dance style music (BRD_o in the ongoing) from the ISMIR 2004 tempo induction contest[1] [9] were included (cf. Table 1). 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]. As with the Bello data set, all annotations have been revised for network training.

Set	# files	# onsets	min/max/mean length [s]
BRD_o	87	5474	10.0 / 10.0 / 10.0
PNP	1	93	13.1 / 13.1 / 13.1
PP	9	489	2.5 / 60.0 / 10.5
NPP	6	212	1.4 / 8.3 / 4.3
MIX	7	271	2.8 / 15.1 / 8.0

Table 1. Statistics of the onset data sets.

For network training, the full set (BRD_o and Bello set) is initially randomly split on the file level into eight disjunctive folds. Next, in an 8-fold cross validation, results for the full set are obtained. Thereby for each fold six subsets are used for training, one for validation, and one for testing. Since the initial weights of the neural nets are randomly distributed, the 8-fold cross validation is repeated 10 times (using the exact same folds) and the means of the output activation functions are used for the final evaluation.

6. RESULTS

In [2] and [5], an onset is reported as correct if it is detected within a 100 ms window (±50 ms) around the annotated ground truth onset position. In [3] a smaller window of $\pm 25ms$ was used for percussive sounds. We therefore decided to report two results for each set, one using a 100 ms window ω_{100} for comparison with results in [2] and [5], and the second using a 50 ms window ω_{50}. All results were obtained with a fixed threshold scaling factor of $\lambda = 50$.

Table 2 shows the results of our BLSTM network approach for each set of onsets in comparison to six other onset detection methods as reported in [2] and [5].The PNP data set consists of 93 onsets from only one audio file of string sounds. As a consequence, the results are not as representative as the others, and can vary a lot, depending on the used parameters, as shown by [5]. The number of onsets of the PP set has changed from originally 489 (used in [2, 5]) to 482 now, due to modifications by its author. The new results are therefore slightly worse (up to max. 1.4%) than the original results but can still compete.

BRD_o & Bello-set	Precision	Recall	$\mathbf{F_1}$**-measure**
BLSTM (ω_{100})	0.945	0.925	0.935
BLSTM (ω_{50})	0.920	0.901	0.911
BLSTM ($comb, \omega_{100}$)	0.938	0.916	0.927
BLSTM ($comb, \omega_{50}$)	0.911	0.890	0.900

Table 3. 8-fold cross validation results for BLSTM on the full data set with 100 ms and 50 ms detection windows (ω). $comb$: all onsets within 30 ms combined.

Table 3 shows the results obtained by cross validation for the full data set. The first two rows reflect the results obtained with the same settings as for the individual Bello sets. It has been shown that two onsets are perceived as one if they are not more than 30 ms apart [12]. Hence we also report results, where all onsets less than 30 ms apart have been combined to a single one. There are 6 605 onsets in the original annotations and 5 861 after combining.

6.1 Discussion

The results show that our algorithm can compete with, and in most cases outperform, a range of existing methods for all types of onsets. However, we must temper this conclusion by adding that we were not able to compare to the latest MIREX participants (e.g. [16]), since the MIREX test data is not publicly available and the authors did not publish results on the Bello data set. Perhaps the most exciting aspect of our approach is that it does not require adaptation to specific onset types to achieve good results. This is an important step towards a universal onset detector.

If a detection window of only 50 ms is chosen our approach even outperforms the reference algorithms in some cases. This shows the excellent temporal precision of the BLSTM onset detector. In our opinion the results given for a detection window of 50 ms with all onsets less than 30 ms apart combined to a single one should be used in the future, as they better reflect the temporal precision of the algorithm and the perception of the human ear.

7. CONCLUSION

We have presented a novel onset detector based on BLSTM-RNN, which – on the Bello onset data set – achieves results

[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/

	PNP			PP			NPP			MIX		
	P	R	F	P	R	F	P	R	F	P	R	F
HFC [2]	0.844	0.817	0.830	0.947	0.941	0.944	**1.000**	0.967	0.983	0.888	0.845	0.866
SD [2]	0.910	0.871	0.890	0.983	0.949	0.966	0.935	0.816	0.871	0.886	0.804	0.843
NLL [2]	**0.968**	0.968	**0.968**	0.968	0.924	0.945	0.980	0.929	0.954	0.889	0.860	0.874
SF [5]	0.938	0.968	0.952	0.981	**0.988**	0.984	0.959	0.975	0.967	0.882	0.882	0.882
NWPD [5]	0.909	0.968	0.938	0.961	0.981	0.971	0.950	0.966	0.958	0.916	0.845	0.879
RCD [5]	0.948	**0.978**	0.963	0.983	0.979	0.981	0.944	0.983	0.963	**0.945**	0.819	0.877
BLSTM (ω_{100})	**0.968**	0.968	**0.968**	**0.987**	0.987	**0.987**	0.991	**0.995**	**0.993**	0.941	**0.897**	**0.918**
BLSTM (ω_{50})	0.918	0.957	0.937	0.955	0.981	0.968	0.982	**0.995**	0.989	0.844	0.865	0.855

Table 2. Results for the Bello data sets PNP, PP, NPP, and MIX. Precision (P), Recall (R), and F_1-measure (F) (as used in [5]). BLSTM with 100 ms and 50 ms detection windows (ω) in comparison to other approaches: high frequency content (HFC), spectral difference (SD), negative log-likelihood (NLL), spectral flux (SF), normalised weighted phase deviation (NWPD), and rectified complex domain (RCD).

on par with or better than existing results on the same data (wrt. F_1-measure), regardless of onset type. We have also introduced a new thoroughly annotated data set of onsets in ballroom dance music.

The average improvement on the whole Bello data set, is 1.7% F_1-measure absolute. The improvement was best (3.6% F_1-measure, absolute) for complex music mixes, reflecting the adaptivity of our method to different musical genres. Competitive results are obtained even if the detection window is halved in size (50 ms instead of 100 ms).

In future work we will investigate whether the approach is suitable for identifying the onset type (e. g. instrument type, vocal, etc.) via detectors trained on respective data.

8. ACKNOWLEDGMENT

We would like to thank Juan Bello, Alexandre Lacoste, and Douglas Eck for sharing their annotated onset data sets.

9. REFERENCES

[1] M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, 1993.

[2] 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.

[3] N. Collins. A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. In *Proc. of the AES Convention 118*, pages 28–31, 2005.

[4] N. Collins. Using a pitch detector for onset detection. In *Proc. of ISMIR*, pages 100–106, 2005.

[5] S. Dixon. Onset detection revisited. In *Proc. of DAFx-06, Montreal, Canada*, pages 133–137, Sept. 2006.

[6] C. Duxbury, J. P. Bello, M. Davies, M. Sandler, and M. S. Complex domain onset detection for musical signals. In *Proc. DAFx-03 Workshop*, 2003.

[7] F. Eyben, B. Schuller, and G. Rigoll. Wearable assistance for the ballroom-dance hobbyist - holistic rhythm analysis and dance-style classification. In *Proc. of ICME 2007*, pages 92–95. IEEE, July 2007.

[8] F. Eyben, M. Wöllmer, and B. Schuller. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. of ACII 2009*, pages 576–581. IEEE, September 2009.

[9] 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.

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

[11] S. Hochreiter, Y. Bengio, P. Frasconi and J. Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. *A Field Guide to Dynamical Recurrent Neural Networks* IEEE Press, 2001.

[12] S. Handel. *Listening: an introduction to the perception of auditory events*. MIT Press, 1989.

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

[14] A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Proc. of ICASSP'99*, vol. 6, pages 3089–3092, 1999.

[15] A. Lacoste and D. Eck. Onset detection with artificial neural networks. MIREX, 2005.

[16] A. Röbel. Onset Detection By Means Of Transient Peak Classification In Harmonic Bands. MIREX, 2009

[17] R. Zhou and J. Reiss. Music onset detection combining energy-based and pitch-based approaches. MIREX, 2007.