MULTI-RESOLUTION LINEAR PREDICTION BASED FEATURES FOR AUDIO ONSET DETECTION WITH BIDIRECTIONAL LSTM NEURAL NETWORKS

Erik Marchi[1], *Giacomo Ferroni*[2], *Florian Eyben*[1], *Leonardo Gabrielli*[2], *Stefano Squartini*[2], *Björn Schuller*[3,1]

[1]Machine Intelligence & Signal Processing Group, Technische Universität München, GERMANY
[2]A3LAB, Department of Information Engineering, Università Politecnica delle Marche, ITALY
[3]Department of Computing, Imperial College London, UK

ABSTRACT

A plethora of different onset detection methods have been proposed in the recent years. However, few attempts have been made with respect to widely-applicable approaches in order to achieve superior performances over different types of music and with considerable temporal precision. In this paper, we present a multi-resolution approach based on discrete wavelet transform and linear prediction filtering that improves time resolution and performance of onset detection in different musical scenarios. In our approach, wavelet coefficients and forward prediction errors are combined with auditory spectral features and then 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. We compare results with state-of-the-art methods on a dataset that includes Bello, Glover and ISMIR 2004 Ballroom sets, and we conclude that our approach significantly outperforms existing methods in terms of F-Measure. For pitched non percussive music an absolute improvement of 7.5% is reported.

Index Terms— Audio Onset Detection, Linear Prediction, Discrete Wavelet Transform, Neural Networks, Bidirectional Long-Short Term Memory

1. INTRODUCTION

Audio Onset Detection (AOD) aims to identify the single temporal instant that characterises the beginning of an acoustic event. Automatic detection of events in audio signals is exploited in many audio applications including content delivery, compression, indexing, retrieval [1], automatic music transcription [2, 3, 4, 5, 6], and beat detection [7]. A note onset is the single instant that marks the beginning of the transient. It must not be confused with the attack, which is the time interval during which the amplitude envelope increases, or the transient, which is a generic term for the time interval needed for a note to settle into quasi-stationary conditions, and thus, is characterized by fast time-varying amplitude, phase or spectrum. An onset can be classified into two main categories: *hard* and *soft* onset. The former is characterised by steep attack and abrupt changes (e. g., percussion instruments) that make it simple to detect by analysing the energy, conversely the latter has a smooth attack (e. g., strings or bowed and wind instruments) for which energy-based onset detection has poor performance.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 289021 (ASC-Inclusion). Correspondence should be addressed to erik.marchi@tum.de.

Fig. 1. Basic onset detection block diagram.

1.1. Related work

Several onset detection methods have been proposed in the recent years and they traditionally rely only on spectral and/or phase information. Energy-based approaches [1, 7, 8] show that energy variations are quite reliable in discriminating onset position especially for hard onsets. Other more comprehensive studies attempt to improve soft-onset detection using phase information [1, 8, 9], and combine both energy and phase information to detect any type of onsets [10, 11, 12, 13]. Further studies exploit the multi-resolution analysis [14] getting advantage from the sub-band representation, and apply a psychoacoustics approach [15, 16] to mimic the human perception of loudness. Finally, other methods use the linear prediction error obtaining a new onset detection function [17, 18, 19]. In particular we will compare our proposed method with common approaches such as spectral difference (SD) [1], high frequency content (HFC), spectral flux (SF) [20], and super flux [21] that basically rely on the temporal evolution of the magnitude spectrogram by computing the difference between two consecutive short-time spectra. Furthermore we evaluate other approaches based on auditory spectral features (ASF) [7] and on complex domain (CD) [22] that incorporates magnitude and phase information.

1.2. Contribution

A traditional onset detection work-flow is given in Figure 1: the input audio signal $x[n]$ is preprocessed and suitable features are extracted. The feature vectors are then processed by the onset detection function (ODF) before detecting the actual onsets via peak detection function. In this paper we propose a novel approach that relies on Wavelet Coefficients (WCs), and Forward Prediction Errors (FPEs) envelope to detect the onsets by exploiting the non-stationary property of the onset [17]. The novel coefficients combined with auditory spectral features [7] are used as input for a Bidirectional Long Short-Term Memory (BLSTM) recurrent neural network [23] which acts as a reduction operator leading to the onset position. We show that our novel approach significantly outperforms existing methods.

After detailing the multi-resolution and linear prediction based coefficients in Section 2, we describe the LSTM Neural Networks in Section 3. Section 4 describes the experiments conducted, before some conclusions are drawn and the relation to prior work is discussed.

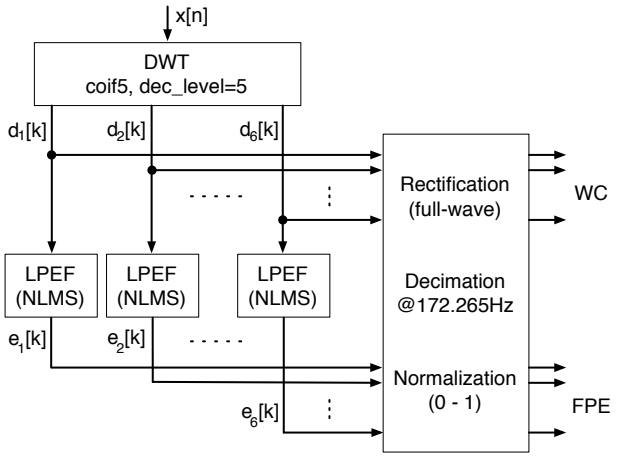

Fig. 2. Features extraction detailed block diagram.

2. MULTI-RESOLUTION LINEAR PREDICTION BASED FEATURES

The feature extraction process is based on the Discrete Wavelet Transformation (DWT) of the input audio signal leading to a sub-band multi-resolution representation. WCs are then processed by Linear Prediction Error Filters (LPEFs) in order to extract the FPEs. The latter, together with the WCs plus their first *average* derivatives constitute the novel feature set. The general block scheme is depicted in Figure 2.

2.1. Discrete Wavelet Transform

The input signal is decomposed in sub-bands applying a multi-resolution analysis computed by a dyadic filter bank as in [19]. We chose 5 decomposition levels obtaining 6 sub-bands.

Given their property of biorthogonality, we adopted Coiflets as discrete wavelets. Coiflets provide nearly linear phase, and a high number of vanishing points that increase convergence efficiency of the LMS algorithm [24].

In order to avoid misalignment among the sub-band signals caused by the asymmetric tree structure of filter bank, the wavelet output coefficients require a delay compensation.

Considering that Coiflets of order 5 have an impulse response length $N = 30$, we can precisely evaluate each band delay after downsampling via the following equation,

$$D_j = \left\lfloor \sum_{j=1}^{J} \frac{\frac{N-1}{2}}{2^j} \right\rfloor \tag{1}$$

where $j = 1$ is the highest band while $j = J$ is the lowest band and $\lfloor . \rfloor$ indicates the floor operation.

2.2. Linear Prediction Filter

Each sub-band signal is processed by an LPEF whose coefficients are updated with each input sample by a modified version of the Normalized LMS (NLMS) algorithm [24] described below. The NLMS approach is chosen for its suitability to signals with large energy variations, such as music signals.

In order to detect onsets by observing the prediction error, the step-size value for the j-th band, μ_j, is crucial,

$$\mu_j = \frac{\mu'}{|\mathbf{u}_j[k]|^2 + c} \tag{2}$$

where $0 < \mu' < 2$, c is a small constant to avoid division by zero, $\mathbf{u}_j[k] = (d_j[k-1], \ldots, d_j[k-p])^T$ represent the previous p input samples and $|.|$ acts as estimate of the signal energy, which varies in time, making the step-size varying as well. However, if the convergence of the filter coefficients is too fast, the increment of the prediction error envelope at note boundary may became less evident, thus, a large value of the step-size is not desired for our task. Hence, we applied a modified step-size that considers silence regions of some kind of music (e. g., pitched non percussive), as reported in [18]:

$$\mu = \min\left(\frac{A}{rms[k] \cdot p}, \frac{1}{|\mathbf{u}_j[k]|^2}, 1000\right) \tag{3}$$

where $rms[k]$ is the root mean-square value of samples in a 20 ms window just after the k-th sample of $d_j[k]$. The constant A is empirically set to 0.5. The second term in the minimum operation ensures the convergence while the third term prevents the step-size from getting too large when the signal energy becomes very small. The filter order p assumes different values depending on the sub-band sample frequency. For the highest band $p_{max} = 24$ while for the lowest two bands $p_{min} = 16$. These parameter values are defined as result of several preliminary evaluations.

2.3. Feature refinement

WCs and FPEs of each band are used as features but a further processing is required in order to use them with the neural network. Due to the multi-resolution nature of the wavelet transformation, each sub-band signal has different resolution. Thus, we chose a suitable sample frequency in order to avoid non-integer decimation: $F_s = \frac{44100}{2^8} = 172.265$ Hz leading to the time resolution equal to $T_r = 5.8$ ms.
WCs and FPEs are rectified by a full-wave rectifier function and decimated to obtain the desired time resolution. Furthermore, to obtain a better functioning of the neural network, they are normalised according with the min-max normalisation, $\bar{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$. Figure 2 shows the complete signal flow.

In order to extract information on time evolution of preceding features, their first order average positive differences are added applying the function $H(x) = \frac{x + |x|}{2}$ to the difference among the n-th sample and the average value of preceding 5 samples corresponding to 29 ms,

$$\begin{aligned} WC^{+}_{avg} &= WC_{n,j} - avg\{WC_{(n-5):(n-1),j}\} \\ FPE^{+}_{avg} &= FPE_{n,j} - avg\{FPE_{(n-5):(n-1),j}\} \end{aligned} \tag{4}$$

with n being the sample index and j the band index. The length of the average window arise from the conducted evaluations.

Summing up, the presented feature set – referred as WC-LPE – is composed of:

- WCs obtained by the filterbank and their corresponding first order positive differences (WC^{+}_{avg}), resulting in 12 features.
- $FPEs$ of each sub-band and their corresponding first order average positive differences (FPE^{+}_{avg}), resulting in 12 features.

2.4. Auditory spectral features

In order to explore the efficacy of our novel feature set, we conducted further experiments by merging the proposed features with Auditory Spectral Features (ASF) [7]. ASF are computed by applying two Short Time Fourier Transform (STFT) using different frame

lengths 23 ms and 46 ms. Each STFT yields the power spectrogram which is converted to the Mel-Frequency scale using a filter-bank with 40 triangular filters obtaining the Mel spectrograms $M_{23}(n,m)$ and $M_{46}(n,m)$. Finally, to match the human perception of loudness, a logarithmic representation is chosen:

$$M_{log}^{23|46}(n,m) = log(M_{23|46}(n,m) + 1.0) \quad (5)$$

In addition the positive first order differences $D_{23}^{+}(n,m)$ and $D_{46}^{+}(n,m)$ are calculated from each Mel spectrogram following:

$$D_{23|46}^{+}(n,m) = M_{log}^{23|46}(n,m) - M_{log}^{23|46}(n-1,m) \quad (6)$$

Mel spectrograms plus first order differences are computed using a frame length of 23 ms are referred as ASF_{23} while for a frame length of 46 ms we refer to ASF_{46}. ASF indicates the combination of the two feature sets. In order to combine WC-LPE with ASF, we adapted the original ASF set to F_s (cf. Section 2.3).

3. BLSTM NEURAL NETWORK AND PEAK DETECTION

A suitable type of network for our purpose is a Bidirectional Recurrent Neural Network (BRNN) with LSTM units instead of usual non-linear one. BLSTM networks have been already applied for onset and beat detection tasks [7] with remarkable performance.
We conducted several preliminary evaluations to find the best network layout by varying the number of hidden layers and their size (i.e. number of LSTM units for each layer). The best network layout for WC-LPE feature set has four hidden layers (two for each direction) with 40 LSTM units each, while for all the others combination tests, the best network has six hidden layers in total (three for each direction) with 20 LSTM units, each, as well as using the ASF set alone. Supervised learning with early stopping was applied for training the network. Network weights are recursively updated by standard gradient descent with backpropagation of the output error. The gradient descent algorithm requires the network weights to be initialised with non zero values; thus we initialise the weights with a random Gaussian distribution with mean 0 and standard deviation 0.1. The output layer of both the networks has one unit and its output activation function lies between 0 and 1 representing the probability for the class 'onset'. Thus, the trained network is able to classify each sample as 'onset'. Samples containing the onset are identified by processing the output unit function: higher output activation function values indicate an high probability that the sample is an onset. An adaptive threshold technique has been implemented before peak picking in order to find the best threshold for each song dealing with the network output. Thus, a threshold θ is computed per song in accordance with the median of the activation function, fixing the range from $\theta_{min} = 0.1$ to $\theta_{max} = 0.3$:

$$\theta' = \beta \cdot median\{a_0(1), ..., a_0(N)\} \quad (7)$$

$$\theta = min(max(0.1, \theta'), 0.3) \quad (8)$$

where $a_0(n)$ is the output activation function of the BLSTM network (sample $n = 1...N$) and the scalar value β is chosen to maximise the F-measure on the validation set. The final onset function $o_o(n)$ contains only the activation values greater than this threshold.

4. EXPERIMENTS AND RESULTS

The aim of our experiments is to evaluate first the performance of ASF and the novel features sets in isolation. Then, we evaluate the combination of them.

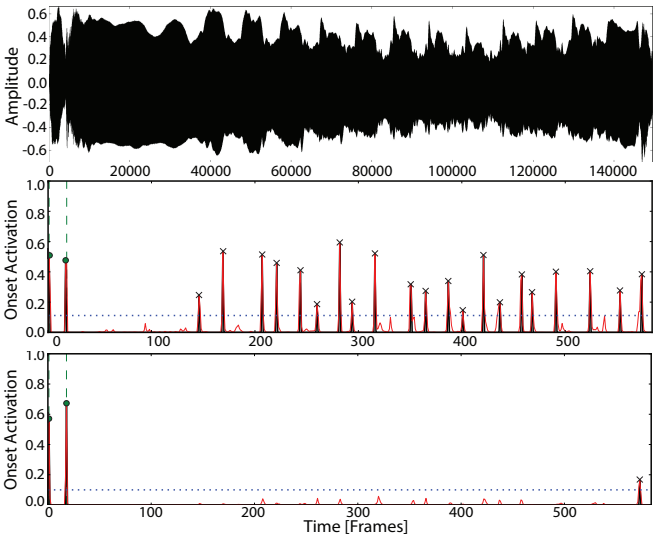

Fig. 3. *Top*: PNP song excerpt *shenai-C.wav*. *Middle*: network output using ASF. *Bottom*: network output using WC-LPE + ASF_{23+46} – only one false positive is present. Correct detection (green circle), false positive (cross), ground truth onsets (vertical dotted line), song threshold (horizontal dotted line), network output before (red line) and after (black line) thresholding and peak picking.

4.1. Evaluation Data Set

The dataset consists of 182 audio excerpts. It was created taking Bello's dataset [1], the dataset used by Glover et al. in [25] and the publicly available labelled excerpts from the ISMIR 2004 Ballroom set[1]. The final set was processed as monaural signals sampled at 44.1 kHz. It is composed of different categories of music[2] pitched percussive (PP e. g., piano), pitched non-percussive (PNP e. g., bowed strings), non-pitched percussive (NPP e.g., drums), complex mixture (MIX e.g., pop music) for a total amount of 7329 onsets. In details the data set contains 6 025 onsets of MIX music, 638 of PP, 306 of PNP, and 360 onsets of NPP music.

4.2. Experimental Setup

In all experiments we evaluate by means of 8-fold cross-validation. Common metrics have been used to evaluate the performance: Precision, Recall and F-measure. The results are reported using a tolerance window of $\pm$25 ms and $\pm$50 ms. First, we evaluate our approach more deeply by applying only WC-LPE features. Then, we incrementally add auditory spectral features. In order to have a more comprehensive comparison with existing approaches we conducted a second group of experiments – again on the full dataset – adopting the same evaluation procedure reported in [9]. Thus, while in the first group of experiments we used an evaluation method that does not contemplate double detections for single target or single detection for double close targets within the tolerance window, in the second group of experiments we applied a more restrictive evaluation [9] that considers one false positive for double detections and one false negative for single detection of two close targets within the tolerance window. Additionally, we only show results with a tolerance window of $\pm$25 ms.

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

[2] Bello and Glover datasets specify the music categories. The ISMIR 2004 Ballroom has been included in the MIX set being polyphonic, multi-instrument, contemporary ballroom music.

Feature Sets	Full dataset			Type subset (F-measure)			
	Precision	**Recall**	**F-measure**	**PP**	**NPP**	**PNP**	**MIX**
ASF (ω_{100})	0.908	0.949	0.928	0.978	0.969	0.831	0.926
ASF (ω_{50})	0.872	0.926	0.898	0.968	0.958	0.804	0.893
WC-LPE (ω_{100})	0.939	0.928	**0.933**	0.976	**0.974**	**0.844**	**0.931**
WC-LPE (ω_{50})	0.897	0.887	0.892	0.961	0.965	0.812	0.885
WC-LPE + ASF_{23} (ω_{100})	0.933	0.942	0.937*	**0.984**	0.972	0.868	0.934
WC-LPE + ASF_{23} (ω_{50})	0.892	0.910	0.901	**0.973**	0.948	0.835	0.894
WC-LPE + ASF_{46} (ω_{100})	0.920	0.952	0.936*	0.980	0.968	0.853	0.934
WC-LPE + ASF_{46} (ω_{50})	0.897	0.922	0.900	0.968	0.966	0.824	0.894
WC-LPE + ASF (ω_{100})	0.952	0.932	**0.942***	0.981	**0.986**	**0.906***	**0.937***
WC-LPE + ASF (ω_{50})	0.923	0.897	0.910	0.969	**0.979**	**0.877***	**0.901**

Table 1. Results for the entire evaluation data set (Full dataset) and for different types subset PNP, PP, NPP, and MIX. Precision (P), Recall (R), and F1-measure (F). BLSTM with tolerance windows of $\pm$50 ms (i.e. ω_{100}) and of $\pm$25 ms (i.e. ω_{50}) using different feature sets: Auditory Spectral Features (ASF) [7], Wavelet Coefficients and Linear Prediction Errors (WC-LPE), WC-LPE plus Mel-spectrum features and first order differences (WC-LPE + $ASF_{23|46}$) and combined feature set (WC-LPE + ASF). * indicates significant improvement (one-tailed z-test, $p<0.05$).

4.3. Results

Table 1 reports onset detection performances for different feature sets using two different tolerance windows within which onsets are correctly detected. We evaluated the different feature sets on the entire dataset and on the four different music types. ASF shows good performances both on the entire dataset and on each type of music with the exception of the PNP set because of the smooth note *attack* present in pitched non percussive music. The WC-LPE feature set alone gives extremely competitive performance and it outperforms ASF in all cases except PP. For PNP type of music, it shows an absolute improvement of 1.3%. The proposed feature set interestingly shows good performances also with respect to the low dimensionality of the feature space: 4 134 features/sec (cf. Section 2.3) against the 27 562 features/sec of auditory spectral approach. Then, we incrementally added auditory spectral features by adding only spectral feature obtained with 23 ms (ASF_{23}) or 46 ms (ASF_{26}) window length and an increase in performance can be observed in Table 1. Finally, we added the full auditory spectral feature set and we obtained better performance in every type of music and on the entire dataset as well (with respect to F-measure). We observe an absolute improvement of 7.5% in the PNP set. Figure 3 shows the improved robustness to false positives in PNP music provided by the proposed combined WC-LPE + ASF feature set with respect to ASF only. In particular the song excerpt (cf. top Figure 3) shows one note sustained with vibrato and tremolo technique which are prone to cause false positives (cf. middle Figure 3). By adding the WC-LPE features the false positives rate vastly decreases (cf. bottom Figure 3). This valuable behavior arises from the ability of LPEF-based feature to deal with: (1) soft onset, which does not generally carry substantial energy changes; (2) different playing techniques (e.g., tremolo), that may cause several false positives due to variations in the intensity of the signal. As an overall evaluation on the full dataset, Figure 4 shows the comparison between state-of-the-art methods and our proposed approach in terms of F-Measure. A significant improvement (one-tailed z-test [27], $p<0.05$) of 1.2% absolute is observed. Note that results are computed with the more restrictive evaluation approach as described in Section 4.2.

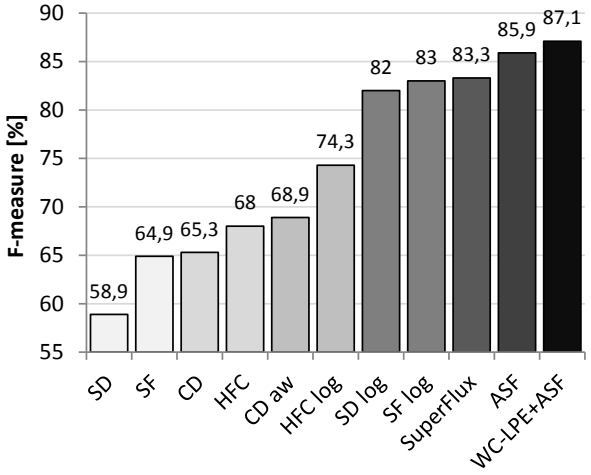

Fig. 4. Comparison with existing methods [9] on the full dataset. Reported approaches are: Complex Domain (CD) [22], Spectral Distance (SD) [1], High Frequency Content (HFC), Spectral Flux (SF) [20], SuperFlux [21]. 'log' indicates log filtering [9]. 'aw' indicates adaptive whitening algorithm [26].

5. CONCLUSION

We have presented a novel multi-resolution linear prediction based approach for audio onset detection, which – on the adopted dataset – achieves better results than existing methods on the same data (with respect to F-measure), regardless of onset type. The absolute improvement on the whole dataset with a more restrictive evaluation is 1.2%. We proposed a new robust feature set that can be efficiently processed by BLSTM. Results corroborate common wisdom that the linear prediction error is carrying relevant information for the onset detection task. In absence of onsets the linear prediction error converges to zero; besides, in non-stationary conditions, such as in correspondence of an onset, the prediction error promptly increases. Hence, our approach has a more robust and efficient soft-onset detection rather than common approaches that simply consider signal energy variations. In fact, we obtained a vast performance improvement in pitched non percussive music, where soft-onset are mainly present since the PNP category includes wind and stringed instruments (e. g., violins, flutes). A preliminary system version was evaluated at MIREX 2013 [28]. Different improvements (cf. Sec. 2.3) led to the presented method. In the future we will investigate the integration of backward prediction errors within the feature set to improve the onset localization accuracy.

6. REFERENCES

[1] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 1035–1047, 2005.

[2] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, pp. 1–28, 2013.

[3] E. Benetos and S. Dixon, "Polyphonic music transcription using note onset and offset detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 37–40.

[4] F. Canadas-Quesada, F. Rodriguez-Serrano, P. Vera-Candeas, N. Ruiz Reyes, and J. Carabias-Orti, "Improving multiple-f0 estimation by onset detection for polyphonic music transcription," in *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*, 2010, pp. 7–12.

[5] C. vd Boogaart and R. Lienhart, "Note onset detection for the transcription of polyphonic piano music," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, 2009, pp. 446–449.

[6] N. Maddage, "Automatic structure detection for popular music," *Multimedia, IEEE*, vol. 13, no. 1, pp. 65–77, 2006.

[7] F. Eyben, S. Böck, B. Schuller, and A. Graves, "Universal onset detection with bidirectional long short-term memory neural networks.," in *ISMIR*, 2010, pp. 589–594.

[8] D. Simon, "Onset detection revisited," in *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06)*, Montreal, Quebec, Canada, Sept. 18–20, 2006, pp. 133–137, `http://www.dafx.ca/proceedings/papers/p_133.pdf`.

[9] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods.," in *Proc. of the International Society for Music Information Retrival Conference*, Porto, Portugal, Oct. 8–12 2012, pp. 49–54.

[10] A. Holzapfel, Y. Stylianou, A. Gedik, and B. Bozkurt, "Three dimensions of pitched instrument onset detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1517–1527, 2010.

[11] R. Z., M. Mattavelli, and G. Zoia, "Music onset detection based on resonator time frequency image," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1685–1695, 2008.

[12] W.C. Lee, Y. Shiu, and C.C. Kuo, "Musical onset detection with joint phase and energy features," in *Multimedia and Expo, 2007 IEEE International Conference on*, 2007, pp. 184–187.

[13] J. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *Signal Processing Letters, IEEE*, vol. 11, no. 6, pp. 553–556, 2004.

[14] C. Duxbury, J. Bello, M. Sandler, and M. Davies, "A comparison between fixed and multiresolution analysis for onset detection in musical signals," in *the 7th Conf. on Digital Audio Effects. Naples, Italy*, 2004.

[15] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, 1999, vol. 6, pp. 3089–3092 vol.6.

[16] B. Thoshkahna and K. Ramakrishnan, "A psychoacoustics based sound onset detection algorithm for polyphonic audio," in *Signal Processing, 2008. ICSP 2008. 9th International Conference on*, 2008, pp. 1424–1427.

[17] W.C. L. and C.C. Kuo, "Musical onset detection based on adaptive linear prediction," in *Multimedia and Expo, 2006 IEEE International Conference on*, 2006, pp. 957–960.

[18] W.C. L. and C.C. Kuo, "Improved linear prediction technique for musical onset detection," in *Intelligent Information Hiding and Multimedia Signal Processing, 2006. IIH-MSP '06. International Conference on*, 2006, pp. 533–536.

[19] L. Gabrielli, F. Piazza, and S. Squartini, "Adaptive linear prediction filtering in dwt domain for real-time musical onset detection," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 650204, 2011.

[20] P. Masri, *Computer modelling of sound for transformation and synthesis of musical signals.*, Ph.D. thesis, University of Bristol, 1996.

[21] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx-13*, Maynooth, Ireland, 2013.

[22] C. Duxbury, J. Bello, M. Davies, and M. Sandler, "Complex domain onset detection for musical signals," in *Proc. Digital Audio Effects Workshop (DAFx)*, 2003.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] S. Haykin, *Adaptive Filter Theory, 4/e*, Pearson Education India, 2005.

[25] J. Glover, V. Lazzarini, and J. Timoney, "Real-time detection of musical onsets with linear prediction and sinusoidal modeling," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–13, 2011.

[26] D. Stowell and M. Plumbley, "Adaptive whitening for improved real-time audio onset detection," in *Proceedings of the International Computer Music Conference (ICMC'07)*, 2007, vol. 18.

[27] M. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007, pp. 623–632.

[28] G. Ferroni, E. Marchi, F. Eyben, L. Gabrielli, S. Squartini, and B. Schuller, "Onset Detection Exploiting Adaptive Linear Prediction Filtering in DWT Domain with Bidirectional Long Short-Term Memory Neural Networks," in *Proceedings Annual Meeting of the MIREX 2013 community as part of the 14th International Conference on Music Information Retrieval*, Curitiba, Brazil, November 2013, ISMIR, ISMIR, 4 pages.