ONSET DETECTION EXPLOITING ADAPTIVE LINEAR PREDICTION FILTERING IN DWT DOMAIN WITH BIDIRECTIONAL LONG SHORT-TERM MEMORY NEURAL NETWORKS

Giacomo Ferroni[2], Erik Marchi[1], 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
{erik.marchi|eyben|schuller}@tum.de
giaferroni@gmail.com,{l.gabrielli|s.squartini}@univpm.it

ABSTRACT

The following short paper presents an experimental algorithm for onset detection which applies multi-resolution and auditory spectral features to Bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks. The proposed algorithm exploits multi-resolution time-frequency features via the discrete wavelet transformation to decompose the input audio signal into sub-bands. Each sub-band is processed by a linear prediction error filters, obtaining the prediction error. The prediction errors together with the wavelet coefficients, their temporal differences and the well-known auditory spectral features are used as input units for the supervised learning. The algorithm has been tested against the MIREX 2013 onset dataset.

1. ALGORITHM DESCRIPTION

The main challenge of this task lies in the audio input representation which should provide optimal features for the onset detection. Our approach is based on linear prediction filtering in the wavelet domain as in [3]. The main difference with the cited approach lies in the application of a bidirectional recurrent neural network with Long Short-Term Memory units (LSTM [6]) to obtain an Onset Detection Function (ODF).

Audio signals are generally composed by stationary or quasi-stationary parts and by *transients* which, conversely, violates the stationary condition playing an important role in the perception of music for humans and consequently in the onsets detection. Indeed a signal modelled by a linear prediction filter gives a prediction error signal tending to zero during the stationary parts but, at the note boundary, the prediction error envelope increases. Consequently, the onset can be located by analysing the prediction error signal. Wavelet analysis is applied to obtain a subbands signal representation and for the fast convergence speed of adaptive prediction filters approach in the transformed domain [9].

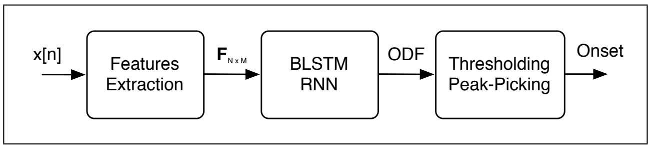

Figure 1. General algorithm block-scheme. $x[n]$ represents the discrete input audio file, $\mathbf{F}_{NxM}$ indicates the features matrix and ODF is the Onset Detection Function.

In order to obtain a suitable audio input representation, the input signal $x[n]$ is firstly decomposed in different subbands using a dyadic filter bank based on wavelet filter coefficients. Each band is, thus, modelled by a Linear Prediction Error Filter (LPEF) and its coefficients are updated by a modified version of a well-know adaptive technique: Normalized LMS (NMLS). We preferred an adaptive approach instead of optimal solution search because the filter's coefficients are continuously updated so that non-stationary parts (i.e., note boundary) produce a significant increment of prediction error envelope. Due to different lengths of the wavelet coefficients (i.e., filter bank output signals) and prediction errors (i.e., LPEF output signals) and in order to use them as neural network inputs, they are re-sampled at a predetermined rate and normalized. Furthermore their first order positive differences are computed. Finally, in order to obtain better performance, a subset of auditory spectral features [2] are added to preceding sets leading to the features matrix $\mathbf{F}_{NxM}$ where M is the number of features and N is the "frame" index.

This matrix is, thus, used as input of a bidirectional recurrent neural network with Long Short-Term Memory units (BLSTM). Network acts as a reduction operator leading to the ODF.

Finally a thresholding and peak-picking algorithm is applied to ODF in order to identify the correct onset positions. Algorithm block-scheme is showed in Figure 1 and block details are described in the following sections.

1.1 Feature Extraction

Discrete input audio files, mono sampled at $F_s = 44.1$kHz, have been used for our experiments.

1.1.1 Discrete Wavelet Transformation

The input file is decomposed in sub-bands applying a multi-resolution analysis computed by a dyadic filter bank (cf. Figure 2) as in [3].

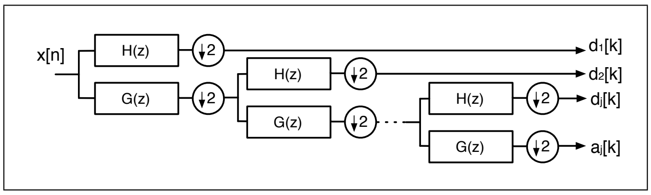

Figure 2. Scheme of a dyadic filter band with j decomposition level.

Concerning the filters impulse response, the Coiflets functions have been used due to their properties: bi-orthogonality, which gives linear or nearly linear phase; high number of vanishing points, that increases convergence properties of the LMS algorithm. We chose $J = 8$ decomposition level obtaining 9 subbands. Due to several experiments we decided to discard the lowest band (from 0 Hz to 86 Hz) because it carries noise and less information that degrades the overall performance.

The alignment among bands is necessary for our task since we need to avoid misalignment which compromises the algorithm precision. Indeed, the wavelet output coefficients require a delay compensation due to asymmetric tree structure of filter bank. The down-sampling must be taken into account during delays evaluation. The chosen wavelet function has nearly linear phase property, thus, the group delay can be approximated to the one of a linear phase filter with $(N-1)/2$ samples, where N is the impulse response length.

Considering that Coiflet wavelet of order 5 has an impulse response length $N = 30$, we are able to precisely evaluate each band delay with regard to the applied down-sampling.

The highest band is delayed by: $\lfloor D_1 = [(N-1)/2]/2\rfloor$ samples, the band immediately below by: $\lfloor D_2 = D_1 + [(N-1)/2]/4\rfloor$ and so on. In general:

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

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

1.1.2 Linear Prediction Error Filter

Each sub-band signal is fed to a LPEF whose coefficients are updated with each input sample by a modified version of NLMS algorithm that is explained below.

Referring to Figure 2, $d_j[k]$ is the $j\text{-}th$ input of the LPEF and assume that $e_j[k]$ is the $j\text{-}th$ prediction error signal. For each j, common LMS iteration consist of:

$$\begin{aligned} y_j[k] &= \mathbf{w}_j^T[k]\mathbf{u}_j[k] \\ e_j[k] &= d_j[k] - y_j[k] \\ \mathbf{w}_j[k+1] &= \mathbf{w}_j[k] + \mu e_j[k]\mathbf{u}_j[n] \end{aligned} \qquad (2)$$

where $\mathbf{u}_j[k] = (d_j[k-1], \ldots, d_j[k-p])^T$ represent the previous p input samples, $\mathbf{w}_j[k] = (a_1, \ldots, a_p)^T$ refer to the FIR filter coefficients, p is the predictor order and μ is the step-size.

In order to detect onsets by observing prediction error, the choice of μ is crucial. Generally we desire that filters coefficients converge as fast as possible to the optimal solution. Dealing with musical signals, the NLMS approach is often chosen for its suitability to signals with large energy variations, such as music:

$$\mu_j = \frac{\mu'}{|\mathbf{u}[k]|^2 + c} \qquad (3)$$

where $0 < \mu' < 2$, c is a small constant to avoid division by zero 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, increment of the prediction error envelope at note boundary may became less evident, thus, a large value of the step-size (as in (3)) is not desired for our task.

The best choice concerning the step-size is reported in [7]:

$$\mu = \min\left(\frac{A}{rms[k]\cdot p}, \frac{1}{|\mathbf{u}[k]|^2}, 100\right) \qquad (4)$$

where $rms[k]$ is the root mean-square value of samples in a 20ms window just after the $k\text{-}th$ sample of $d_j[k]$. Constant A is empirically set to 0.5. The second term in the minimum operation ensures the convergence 3, while the third term prevents the step-size from getting too large when the signal energy becomes very small. This version of NLMS adaptive approach is chosen in our algorithm.

Finally, we used different values for the filter order p for each sub-band. The lowest signal band $d_8[k]$ is fed to a LPEF of order $p_{min} = 10$ while an order $p_{max} = 24$ is used with the highest signal band $d_1[k]$. The following rule is applied:

$$p_j = 10 + 2\cdot(J-j) \qquad j = 1, \ldots, J = 8 \qquad (5)$$

1.1.3 Features refinement

Wavelet Coefficients (WCs) and Prediction Errors (PEs) 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 (i.e., different length respect to other bands). We adopted a suitable time resolution for our task: $T_{res} = 10$ ms, thus, WCs and PEs have been rectified by a full-wave rectifier function and re-sampled to obtain the desired time resolution.

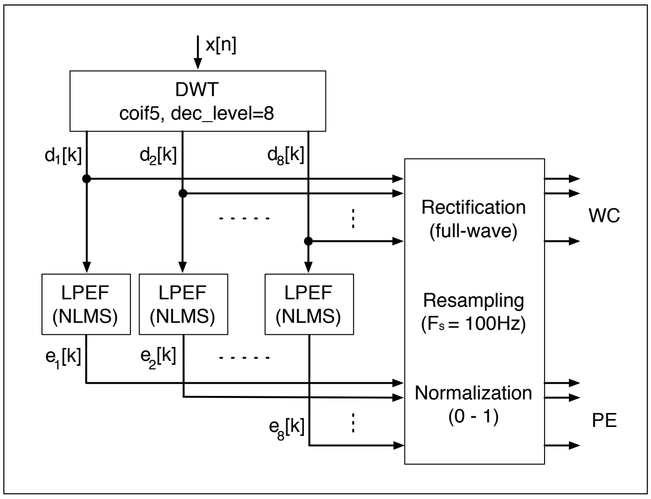

Figure 3. Features extraction details.

Furthermore, to obtain a better functioning of the neural network, they were normalised by a min-max normalisation. Figure 3 shows the complete signal flow.

To exploit the information brought about bt the time evolution of preceding features, the first order positive differences are added applying the function $H(x) = \frac{x-|x|}{2}$.

$$\begin{aligned} WC^{+}_{n,j} &= WC_{n,j} - WC_{n-1,j} \\ PE^{+}_{n,j} &= PE_{n,j} - PE_{n-1,j} \end{aligned} \tag{6}$$

with n being the frame index and j the band index

1.2 BLSTM Neural Network

The best neural network for our purpose is a bidirectional RNN with LSTM units instead of usual non-linear units. As inputs we used 112 features per frame, composed by:

- The prediction errors of each sub-band (PE) and their corresponding first order positive differences (PE^{+}), resulting in 16 features.

- The wavelet coefficients (WC) obtained by the filter bank and their corresponding first order positive differences (WC^{+}), resulting in 16 features.

Plus a subset of auditory spectral features [2]:

- Mel-spectrogram ($M^{log}_{46}(n,m)$) computed with window size of 46.4 ms and its first order positive differences ($D^{+}_{46}(n,m)$), resulting in 80 features.

The network has four hidden layers in total (two for each direction) with 10 LSTM units each.

The output layer has one unit and its output activation function lies between 0 and 1. It represents the probability for the class 'onset' and allows the use of the cross entropy error criterion to train the network [5].

1.2.1 Network Training and Dataset

Supervised learning with early stopping was adopted to train the network. The dataset consists of 199 audio excerpts. It was created taking Bello's dataset [1], the dataset used by Glover et al. in [4], audio files used by Leveau et al. in [8] and some excerpts from ISMIR 2004 Ballroom set [1] .

The final set was processed as monaural signals sampled at 44.1 kHz. It is composed by 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) and others sound (OTHER is composed by ISMIR 2004 Ballroom dataset) for a total amount of 7989 onsets.

Presenting each audio sequence frame by frame to the network, its weights are recursively updated by standard gradient descent with back-propagation of the output error. 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 Thresholding and Peak Picking

The network obtained after training can classify each frame as 'onset' and consequently as 'non-onset' class. Frames containing the onsets are identified by processing the output unit function. Higher output activation function values indicate an high probability that the frame is an onset-frame.

An adaptive threshold technique has to be implemented before peak picking due to the dependency among the input units, namely: detection function, input signal, short time spectrum, wavelet coefficients and prediction errors.

In order to obtain the best classification for each song, a threshold θ is computed per song in accordance with the mean of the activation function, fixing the range from $\theta_{min} = 0.1$ to $\theta_{max} = 0.3$:

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

$$\theta = min(max(0.1, \theta'), 0.3) \tag{8}$$

where $a_o(n)$ is the output activation function of the BLSTM network (frames $n = 1...N$) and the scalar value β is chosen to maximise the F-measure on the validation set. Its value is fixed to $\beta = 3.7$.
The final onset function $o_o(n)$ contains only the activation values greater than this threshold.

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

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

[2] Bello and Glover datasets specify the music categories. ISMIR 2004 Ballroom dataset does not specify these information and we refer to it as OTHER.

2. RESULTS

The presented onset detector attained good performance in the MIREX 2013 evaluation (cf. gray row in Table 1).

Algorithm	F-measure	Precision	Recall
SB1	0.8727	0.8641	0.8946
ZHZD1	0.8233	0.7858	0.9009
FMESS1	0.8062	0.7770	0.8732
FMEGS1	0.8025	0.7880	0.8593
CF4	0.7345	0.6966	0.8507
CB1	0.6308	0.8506	0.5367
MTB1	0.3785	0.5429	0.3418

Table 1. Results for the MIREX 2013 onset detection evaluation. Only the best results of other participants or groups are shown with the exception for our two submissions.

3. ACKNOWLEDGMENT

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).

4. REFERENCES

[1] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. B. Sandler: "A tutorial on onset detection in music signals," *Speech and Audio Processing, IEEE Transactions*, 13.5:1035–1047, 2005.

[2] F. Eyben, S. Böck, B. Schuller and A. Graves: "Universal onset detection with bidirectional long short-term memory neural networks," *11th International Society for Music Information Retrieval Conference (ISMIR 2010*, 2010.

[3] 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 2011*,2011.

[4] 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*, 1:1–13, 2011.

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

[6] S. Hochreiter and J. Schmidhuber: "Long short-term memory," *Neural Computation*, 9.8:1735-1780, 1997.

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

[8] P. Leveau, L. Daudet: "Methodology and tools for the evaluation of automatic onset detection algorithms in music," *In Proc. Int. Symp. on Music Information Retrieval*, 2004.

[9] E. Nurgun and B. Filiz: "Wavelet transform based adaptive filters: analysis and new results, " *Signal Processing, IEEE Transactions on IEEE*. 44.9:2163–2171, 1996.