Audio onset detection: a wavelet packet based approach with recurrent neural networks

Erik Marchi, Giacomo Ferroni, Florian Eyben, Stefano Squartini, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Marchi, Erik, Giacomo Ferroni, Florian Eyben, Stefano Squartini, and Björn Schuller. 2014. "Audio onset detection: a wavelet packet based approach with recurrent neural networks." In *2014 International Joint Conference on Neural Networks (IJCNN), 6-11 July 2014, Beijing, China*, 3585–91. Piscataway, NJ: IEEE. https://doi.org/10.1109/ijcnn.2014.6889669.

Audio Onset Detection: A Wavelet Packet Based Approach with Recurrent Neural Networks

Erik Marchi, Giacomo Ferroni, Florian Eyben, Stefano Squartini, Björn Schuller

***Abstract*— This paper concerns the exploitation of multi-resolution time-frequency features via Wavelet Packet Transform to improve audio onset detection. In our approach, Wavelet Packet Energy Coefficients (WPEC) and Auditory Spectral Features (ASF) are processed by Bidirectional Long Short-Term Memory (BLSTM) recurrent neural network that yields the onsets location. The combination of the two feature sets, together with the BLSTM based detector, form an advanced energy-based approach that takes advantage from the multi-resolution analysis given by the wavelet decomposition of the audio input signal. The neural network is trained with a large database of onset data covering various genres and onset types. Due to its 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 show a comparison with other types and sizes of recurrent neural networks and we compare results with state-of-the-art methods on the whole onset dataset. We conclude that our approach significantly increase performance in terms of F-measure without any music genres or onset type constraints.**

I. Introduction

Onset detection 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 [1], [2], an onset detector faces the challenge of detecting single events, which do not follow a periodic pattern. Recent onset detection methods [3], [4], [5] have matured to a level where reasonable robustness is obtained for polyphonic music. While several methods have been adopted and tuned on specific kinds of onsets (e.g., pitched or percussive), few attempts have been made in the direction of widely-applicable approaches in order to achieve superior performance over different types of music and with considerable temporal precision.

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 [6], [7], [3] 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 [6], [3], [8], and combine both energy and phase information to detect any type of onsets [9], [10], [11], [12]. Further studies exploit the multi-resolution analysis [13] getting advantage from the sub-band representation, and apply a psychoacoustics approach [14], [15] to mimic the human perception of loudness. Finally, other methods use the linear prediction error obtaining a new onset detection function [16], [17], [18]. In particular we will compare our proposed method with common approaches such as spectral difference (SD) [6], high frequency content (HFC), spectral flux (SF) [19], and super flux [20] 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) [21] that incorporates magnitude and phase information.

In the early 1980s, Morlet and Grossmann introduced the transformation method of decomposing a signal into wavelets coefficients and reconstructing the original signal for the first time. After that, at the end of 1980s, Mallat and Mayer developed a multi-resolution analysis using wavelets. Since this new transformation method was born, the wavelet theory has continuously been developed, and nowadays, the Wavelet Transform is widely used in many different fields: Image processing, Digital watermarking, Audio processing among others. The Wavelet Transform is also exploited in Audio processing and Music Information Retrieval (MIR); specifically it is used to extract audio features, as presented in [22], where Discrete Wavelet Transform Octave Frequency Bands are used to create a beat histogram for musical genre classification. In [23], Wavelet-Packet Transform is applied in the field of speech recognition by outperforming the well-known Mel-Frequency Cepstral Coefficients (MFCC). An other important result in speech/music discrimination was obtained in [24] through Wavelet-based parameters. A further Music Onset Detection approach that uses Wavelet Transform and linear prediction filters is presented in [18].

In this paper we propose a novel approach that relies on Wavelet Packet Energy Coefficients (WPEC) to detect the onsets. This is intrinsically multi-resolution due to the wavelet transformation, whereas the auditory spectral features used in [7] requires two transformations, based on the fixed-resolution STFT, with different window lengths. Thus, proven the high onset detection performance achievable using energy-based approach, we aim to build a novel multi-resolution energy-based features set. The novel coeffi-

Erik Marchi, Florian Eyben and Björn Schuller are with the Machine Intelligence & Signal Processing Group, Technische Universität München, Germany (email: {erik.marchi, eyben, schuller}@tum.de). Giacomo Ferroni and Stefano Squartini are with A3LAB, Department of Information Engineering, Università Politecnica delle Marche, Italy (email: giaferroni@gmail.com, s.squartini@univpm.it). Björn Schuller is also with the Department of Computing, Imperial College London, United Kingdom.

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.

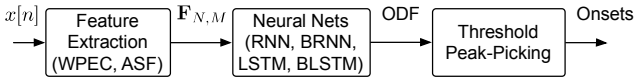

Fig. 1. Common onset detection block diagram.

cients combined with auditory spectral features [7] are then used as input for a Bidirectional Long Short-Term Memory (BLSTM) recurrent neural network [25] which acts as a reduction operator leading to the onset position. Besides showing that our novel approach significantly outperforms existing methods, we also provide a detailed analysis with different types of recurrent neural networks (RNN).

The rest of this paper is structured as follows. A detailed overview of the proposed system is given in Section 2. Section 3 provides a description of the dataset, the experimental set-up and results. Section 4 concludes the paper.

II. System Description

A traditional onset detection work-flow is given in Fig. 1: the input audio signal $x[n]$ is preprocessed and suitable features are extracted. The feature vectors are then processed by the neural network to obtain the onset detection function (ODF) before detecting the actual onsets via peak detection function.

In our approach, the feature extraction process relies on the Discrete Wavelet Packet Transformation (DWPT) of each input signal frame. The sub-bands energies are calculated for each frame in the wavelet domain and additional delta coefficients are employed leading to the Wavelet Packet Energy Coefficient (WPEC) features set. The general block scheme is depicted in Fig. 3.

A. Wavelet Packet Transformation

Discrete Wavelet Packet Transform (DWPT) is a generalisation of the common Discrete Wavelet Transform (DWT). It has emerged as an important signal representation scheme with relevant performance in compression, detection and classification. Discrete Wavelet Transform is very similar, in principle, to the Short-Time Fourier Transform (STFT). While STFT uses a single analysis window, the Wavelet Transform is obtained by dilatations, contractions and shifts of the wavelet function and this method leads to a Multi Resolution Analysis (MRA). It means low time resolution and high frequency resolution at low frequencies and vice-versa at high frequencies. In some music applications, Wavelet Packet Transform can be applied to increase the information available in a part of the frequency axis. DWPT is also an attractive representation because it can be simply implemented with a basic two-channel filter followed by a down-sampling operation. For each level of decomposition, the signal is decomposed into approximation coefficients (output of low-pass filter) and detailed coefficients (output of high-pass filter). While DWT uses the detail coefficients of each level, DWPT decomposes also the detailed coefficients leading to a tree-representation (cf. Fig. 2). Choosing *n* leaves of this decomposition-tree at different depths, we are able to obtain the best time-frequency representation for our task.

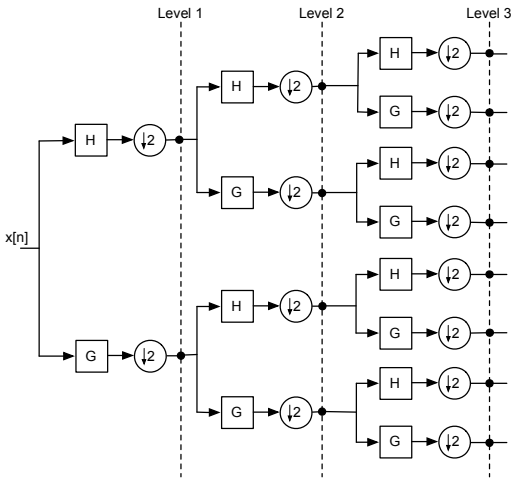

Fig. 2. Example of DWPT implemented by a filter bank.

B. Wavelet Packet Energy Coefficients

The discrete input audio signal $x[n]$ is first segmented into frames of $W = 2048$ samples corresponding to 46ms. The standard Hamming windowing function is afterwards applied to each frame as proposed in [7]: choosing the frame rate $F_f = 100$fps, the hop size h between adjacent windows is equals to F_s/F_f where F_s denotes the sample rate (i.e. $F_s = 44.1$kHz) and they are overlapped of a factor $(W - h)/W$. Each frame is then transformed exploiting the DWPT following the bands division in Table I.

Dec. Level	Level Bandwidth	N. Bands	Frequency Resolution
8	0 ÷ 516 Hz	6	86.13 Hz
7	516 ÷ 1378 Hz	5	172.27 Hz
6	1.38 ÷ 3.45 kHz	6	344.53 Hz
5	3.45 ÷ 5.51 kHz	3	689.06 Hz
4	5.51 ÷ 8.27 kHz	2	1378.13 Hz
3	8.27 ÷ 11 kHz	1	2756.25 Hz
2	11 ÷ 22 kHz	2	5512.50 Hz
Total	**0 ÷ 22**kHz	**25**	-

TABLE I
Frequency band division. Level bandwidth indicates the total bandwidth covered at each level of decomposition.

The sub-bands scheme employed is based on the *critical bandwidth* function derived from the psychoacoustic. The latter aims to characterise human auditory perception and the time-frequency analysis capabilities of the human inner ear [26]. A frequency-place transformation takes place in the cochlea (inner ear), along the basilar membrane. Indeed a sound wave moves the eardrum and the attached ossicular bones, which in turn transfer the vibration to the cochlea that contains the coiled basilar membrane. The travelling waves generate impulses with a relationship between signal frequency and a specific positions of the membrane, along

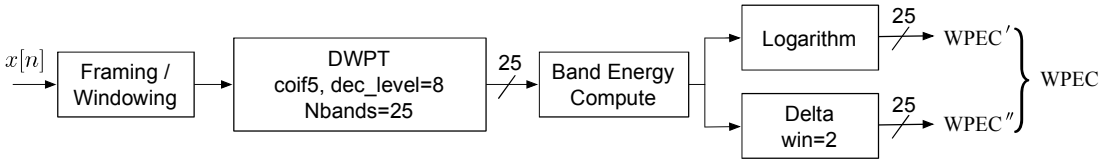

Fig. 3. WPEC general scheme. In the DWPT block, the term *coif5* indicates the wavelet function employed that is the fifth order Coiflets.

which neural receptors are connected. Thus, different neural receptors are effectively able to detect particular frequencies according to their locations. From a signal processing point of view, the cochlea can be seen as a bank of highly overlapping bandpass filters characterised by asymmetric and non-linear magnitude response. Moreover the bandwidth of the filters increases with increasing frequency. The *critical bandwidth* is, thus, a function of frequency that characterizes the cochlear passband filters. The employed DWPT decomposition scheme uses the fifth order Coiflets' wavelet function attempting to mimic this human ear behaviour. In Fig. 4 we report, in a comparative fashion, the plots of the band start frequency in our decomposition scheme and in the *critical bandwidth* function.

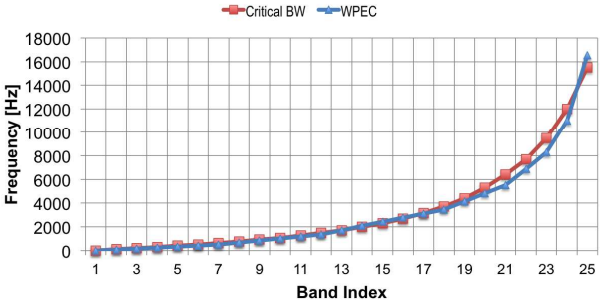

Fig. 4. Band start frequency comparison between critical bandwidth and our wavelet-based decomposition scheme.

The sub-bands are then used to calculated the frame energies vector $E(n,l)$ according to the Eq. (1) where n is the frame index and l is the band index which lies between $j=1$ and $j=25$.

$$E(n,l) = \begin{cases} \sum_k |x_l[k]|^2 + \sum_k |x_{l+1}[k]|^2, & \text{if } l = 1 \\ \sum_k |x_{l-1}[k]|^2 + \sum_k |x_l[k]|^2 + \\ + \sum_k |x_{l+1}[k]|^2, & \text{if } l = 2 \dots 24 \\ \sum_k |x_{l-1}[k]|^2 + \sum_k |x_l[k]|^2, & \text{if } l = 25 \end{cases} \quad (1)$$

Finally, to mimic the human perception of loudness, a logarithmic representation of the energies vectors is chosen (cf. Eq. (2)) and the delta coefficients are extracted applying the half-wave rectifier to the Eq. (3).

$$\text{WPEC}'(n,l) = log(E(n,l) + 1.0) \quad (2)$$

$$\text{WPEC}''(n,l) = \text{WPEC}'(n,l) - \text{WPEC}'(n-2,l) \quad (3)$$

Thus, the final features set is composed by the WPEC′ and WPEC″ consisting of 50 features for each frame. It is indicated simply by WPEC (cf. Fig. 3).

C. Auditory spectral features

In order to have a more exhaustive analysis, further experiments are conducted 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 length 23 ms and 46 ms sampled at a rate of 100 fps. Each STFT yields the power spectrogram which is converted to the Mel-Frequency scale using a filter-bank with 40 triangular filters leading to the Mel spectrograms $M_{23}(n,m)$ and $M_{46}(n,m)$. The logarithmic representation is obtained by:

$$M_{log}^{23|46}(n,m) = log(M_{23|46}(n,m) + 1.0) \quad (4)$$

In addition, the positive first order differences $D_{23}^{+}(n,m)$ and $D_{46}^{+}(n,m)$ are calculated from each Mel spectrogram following the Eq. (5).

$$D_{23|46}^{+}(n,m) = M_{log}^{23|46}(n,m) - M_{log}^{23|46}(n-1,m) \quad (5)$$

Mel spectrograms plus first order differences 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.

D. Neural network and Peak detection

Different kinds of neural networks were analysed in our approach. The most commonly used neural network is the multilayer perceptron (MLP) [27]. This network belongs to the feed forward neural networks (FNNs). A minimum of three layers is needed and all connections feed forward from one layer to the next without any backward connections. To introduce past context to neural network, another technique is to add cyclic connections to FNNs. This backward connections form a sort of memory, which allows input values to persist in the hidden layers and influence the network output in the future. Many different types of cyclic connections were developed in literature [28], [29], [30], [31]. These networks are called recurrent neural networks (RNN). In order to determine the input pattern class affiliation, the future context can be exploited by means of two separated hidden layers. Both of them are connected to the same input and output layer and the input patterns cross the network in both forward and backward directions. These networks, called bidirectional

recurrent neural networks (BRNNs) and they have access to both past and future context in each moment. The main drawback of BRNNs lies in the knowledge of the complete input sequence. It represents a violation of the causality principle leading to disadvantages in on-line applications. Both RNNs and BRNNs exploit standard artificial neurones which generally employ the logistic sigmoid function to their inputs weighted sum. The recurrent connections in RNNs and BRNNs cause the so-called *vanishing gradient problem* [32]. Indeed the input value influence decays or increases exponentially over time, as it cycles through the network via its recurrent connections. By replacing the non-linear units in the hidden layers with the Long Short-Term Memory (LSTM) ones, the vanishing gradient problem is solved. Fig. 5 shows an example of LSTM block.

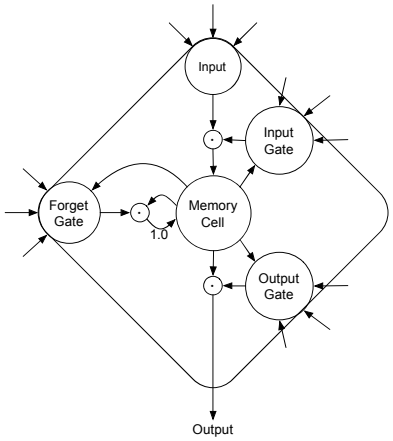

Fig. 5. LSTM block with one memory cell.

It is composed of one or more self connected linear memory cells and three multiplicative gates. The memory cell maintains the internal state for a long time through a constant weighted connection (1.0). The content of the memory cell is controlled by the multiplicative input, output and forget gates. More details can be found in [25], [33].

However, the outcome of a broad number of experiments revealed superior performance in the case of Bidirectional Long Short-Term Memory recurrent neural network [25]. BLSTM network has been already applied for onset and beat detection tasks [7] with remarkable performance.

The proposed set of features (cf. Sect. II-B), WPEC, is firstly used as network input. Following a progressively combination of this set with the ASF (cf. Sect. II-C) one is evaluated in order to compare and merge the two different sets. While WPEC employes 5k features/sec, ASF uses 16k features/sec (i.e., both ASF_{23} and ASF_{46} exploit 8k features/sec). The network has two hidden layers for each direction with 20 LSTM units each and has a single output, where a value of 1 represents an onset frame and a value of 0 a non-onset frame. For network training, supervised learning with early stopping is used. Each audio sequence is presented frame by frame to the network. Standard gradient descend with back propagation of the output errors is used to iteratively update the network weights. The latter are initialized by a random Gaussian distribution with mean 0 and standard deviation 0.1.

The trained network is able to classify each frame into 'onset' or 'non-onset' class (i.e., ideally the output activation value is closest to 1 and 0 respectively). Thresholding and peak detection is therefore applied to the output activations. An adaptive thresholding technique has to be implemented before peak picking because of many onset-frames have the output activation value below the 'standard' threshold for a binary classification (i.e., 0.5). Thus, to obtain the best classification for each song, a threshold θ is computed per song in concordance with the median of the activation function, fixing the range from $\theta_{min} = 0.1$ to $\theta_{max} = 0.3$:

$$\theta' = \lambda \cdot median\{a_o(1), ..., a_o(N)\} \tag{6}$$

$$\theta = min(max(0.1, \theta'), 0.3) \tag{7}$$

where $a_o(n)$ is the output activation function of the BLSTM network (frames $n = 1...N$) and the scalar value λ is chosen to maximise the F_1-measure on the validation set. The final onset detection function $o_o(n)$ contains only the activation values greater than this threshold.

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

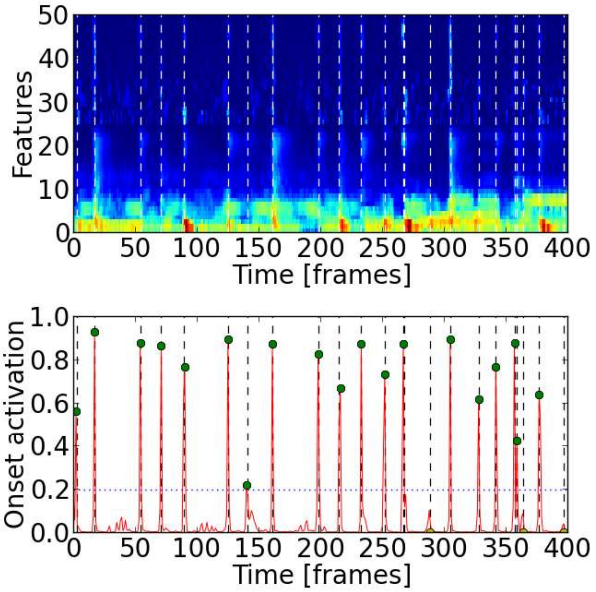

Fig. 6. *Top:* WPEC set with ground-truth onset (vertical dashed lines). *Bottom:* The BLSTM network output before processing (red line) with correctly detected onsets (green dots), erroneous detections (yellow dots), ground-truth onsets (vertical dashed lines) and threshold θ (horizontal dashed line). 4s excerpt from Dido - Here With Me.

WPEC set used with the BLSTM-RNN is depicted in the *Top* of Fig. 6 which refers to an excerpt 4s length of

MIX type. Along the y-axis, coefficients up to 25 represent the logarithmic vector of energies (WPEC$'$) while delta coefficients (WPEC$''$) are represented by coefficients from 26 to 50. Low frequencies energy information are located in the lowest part of both the aforementioned sub-sets. The delta coefficients are very important in the proposed onset detection approach as arose from experiments. The *bottom* of Fig. 6 shown the network output value for each frame (i.e., x-axis) and the song-based threshold. The evaluation algorithm uses the peaks over this threshold to count correct detections (green dots) or erroneous detections (yellow dots).

III. Experiments

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

A. Dataset

The evaluations is computed on a large dataset containing 7329 onsets and distributed in four categories: pitched percussive (PP), non-pitched percussive (NPP), pitched non-percussive (PNP) and complex mixture (MIX). The dimensionality of each categories is reported in Table II.

Type	# files	# onsets
PP	24	638
NPP	22	360
PNP	36	306
MIX	100	6025

TABLE II
NUMBER OF FILES AND ONSETS FORMING THE EMPLOYED DATASET

The dataset is set up taking the Bello's dataset [6], the dataset used by Glover et al. in [34] and some excerpts from the ISMIR 2004 Ballroom set [35]. The whole files are monaural and sampled at 44.1kHz.

B. 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 ± 25 ms and ± 50 ms. First, we evaluate our approach more deeply by applying only WPEC 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 –. 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. We show results with a tolerance window of ± 25 ms and ± 50 ms.

C. Results

Table III reports onset detection performance for different types of neural networks and for different network sizes, using two different tolerance windows within which onsets are correctly detected. The best performance are obtained by using BLSTM recurrent neural network with four hidden layers (two for each direction) composed by 20 LSTM units each. Others types of networks (i.e., RNN, BRNN, LSTM) give good performance however the LSTM block increases the network performance thanks to the ability to classify input patterns, drawing from an extensive part of the past inputs. After a preliminary analysis on the network size and type of network, we evaluated the different feature sets on the entire dataset and on the four different music types.

In Table IV, ASF shows good performance 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 WPEC feature set alone gives competitive performance but it does not outperform ASF. However, the former set exploits less features with respect to ASF, indeed WPEC dimensionality is 5k features per frame while ASF employes 16k features per frame as mentioned above.

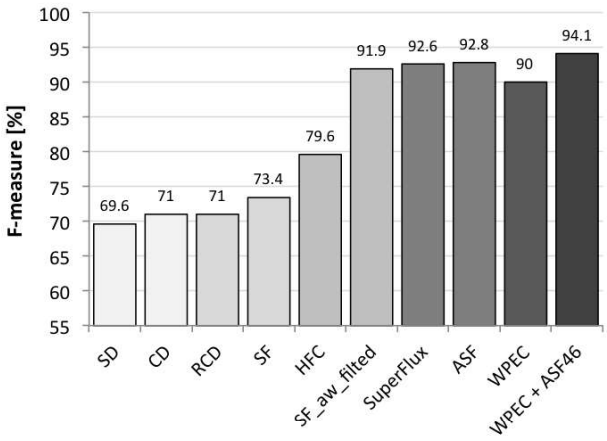

Fig. 7. Comparison with other methods on the full dataset. Reported approaches are: Complex Domain (CD) and Rectified CD [21], High Frequency Content (HFC), Spectral Difference (SD) [6], Spectral Flux (SF) [19], a recently modified SF version [8] and SuperFlux [20]. 'aw' indicates adaptive whitening algorithm [36].

Thus, 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 IV. In the case of WPEC with ASF_{46}, we obtained better performance in every type of music (except pitched percussive) and on the entire dataset as well (with respect to F-measure). The combined set, thus, gives an improvement of overall detection performance with less features per frame with respect to ASF. Indeed, the WPEC + ASF_{46} dimensionality is 13k features per frame, which corresponds to a relative reduction of 18.75%, thus guaranteeing a relevant drop in terms of computational

Net size	RNN			BRNN			LSTM			BLSTM		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
10,10 (ω_{100})	0.867	0.864	0.866	0.884	0.874	0.879	0.877	0.895	0.886	0.926	0.863	**0.894**
10,10 (ω_{50})	0.786	0.803	0.794	0.814	0.815	0.814	0.793	0.839	0.816	0.872	0.801	**0.835**
20,20 (ω_{100})	0.870	0.872	0.871	0.917	0.868	0.892	0.858	0.900	0.878	0.914	0.886	**0.900**
20,20 (ω_{50})	0.878	0.809	0.798	0.860	0.804	0.831	0.773	0.844	0.807	0.857	0.827	**0.842**
10,10,10 (ω_{100})	0.848	0.745	0.793	0.871	0.894	0.882	0.762	0.055	0.102	0.919	0.867	**0.892**
10,10,10 (ω_{50})	0.764	0.694	0.727	0.794	0.836	0.814	0.661	0.051	0.094	0.864	0.807	**0.834**
20,20,20 (ω_{100})	0.856	0.732	0.789	0.913	0.875	0.894	0.869	0.444	0.588	0.904	0.891	**0.898**
20,20,20 (ω_{50})	0.777	0.681	0.726	0.857	0.819	0.837	0.792	0.413	0.543	0.844	0.833	**0.839**

TABLE III

COMPARISON AMONG DIFFERENT NETWORK TYPES AND TOPOLOGIES WITH WPEC FEATURES AS INPUT.

Feature Sets	Full dataset			Type subset (F_1-measure)			
	Precision	**Recall**	**F_1-measure**	**PP**	**NPP**	**PNP**	**MIX**
ASF (ω_{100})	0.920	**0.942**	0.930	0.969	0.969	0.851	0.929
ASF (ω_{50})	0.882	**0.909**	0.895	0.948	0.954	0.807	0.891
WPEC (ω_{100})	0.914	0.886	0.900	0.887	0.951	0.830	0.902
WPEC (ω_{50})	0.857	0.827	0.842	0.853	0.927	0.782	0.839
WPEC + ASF_{23} (ω_{100})	0.944	0.910	0.926	0.957	0.962	0.835	0.926
WPEC + ASF_{23} (ω_{50})	0.912	0.870	0.891	0.938	0.950	0.794	0.887
WPEC + ASF_{46} (ω_{100})	**0.950**	0.933	**0.941**	0.975	**0.982**	**0.880**	**0.939**
WPEC + ASF_{46} (ω_{50})	**0.913**	0.898	**0.906**	0.958	**0.967**	**0.840**	**0.900**
WPEC + ASF (ω_{100})	0.943	0.930	0.936	**0.976**	0.976	0.855	0.934
WPEC + ASF (ω_{50})	0.912	0.899	0.906	**0.967**	0.966	0.822	0.900

TABLE IV

RESULTS FOR THE ENTIRE EVALUATION DATA SET (FULL DATASET) AND FOR DIFFERENT TYPES SUBSET PNP, PP, NPP, AND MIX. PRECISION (P), RECALL (R), AND F_1-MEASURE (F_1). 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 PACKET ENERGY COEFFICIENTS (WPEC), WPEC PLUS MEL-SPECTRUM FEATURES AND FIRST ORDER DIFFERENCES (WPEC + $ASF_{23/46}$), AND COMBINED FEATURE SET (WPEC + ASF).

complexity.

As an overall evaluation on the full dataset, Fig. 7 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 [37], $p<0.05$) of 1.3% absolute is observed. This absolute improvement confirm the effectiveness of the proposed energy-based feature type in the onset detection field and, on the other hand, the benefits provided by the exploitation of multi-resolution time-frequency features via Wavelet Packet Transform.

IV. CONCLUSION

In this contribution, a novel multi-resolution energy based approach for audio onset detection is proposed. The method relies on the multi-resolution analysis of audio data performed by means of Wavelet Packet Transform, and integrates the related features with the auditory spectral features, already used in previous works [7]. The two feature sets are then given as input to a RNN for onset localization: different RNN topologies have been employed and comparatively tested, and the BLSTM resulted to be the most performing one. The overall proposed framework has been then evaluated against several other state of the art methods, showing the best performance with an absolute improvement on the whole dataset of about 1.3%, in terms of F-measure. Moreover, it must be noted that such an improvement is in company with a remarkable reduction in terms of computational complexity.

Future efforts will be targeted to test the proposed approach against a larger dataset as already employed in [20] and to assess its effectiveness by following the evaluation method proposed in [8], which takes double detections for single target onset and single detection for double target onsets into account.

REFERENCES

[1] F. Eyben, B. Schuller, S. Reiter, and G. Rigoll, "Wearable Assistance for the Ballroom-Dance Hobbyist – Holistic Rhythm Analysis and Dance-Style Classification," in *Proceedings 8th IEEE International Conference on Multimedia and Expo, ICME 2007*, Beijing, China, July 2007, IEEE, pp. 92–95, IEEE.

[2] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proceedings 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*, Amsterdam, The Netherlands, September 2009, HUMAINE Association, vol. I, pp. 576–581, IEEE.

[3] S. Dixon, "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.

[4] A. Röbel, "Onset detection by means of transient peak classification in harmonic bands," in *Proceedings of MIREX as part of the 10th International Conference on Music Information Retrieval (ISMIR)*, 2009, p. 2.

[5] R. Zhou and J.D. Reiss, "Music onset detection combining energy-based and pitch-based approaches," *Proc. MIREX Audio Onset Detection Contest*, 2007.

[6] 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 on*, vol. 13, no. 5, pp. 1035–1047, 2005.

[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] 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.

[9] A. Holzapfel, Y. Stylianou, A.C. 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.

[10] Z. Ruohua, 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.

[11] L. Wan-Chi, Yu S., and C.-C.J. Kuo, "Musical onset detection with joint phase and energy features," in *Multimedia and Expo, 2007 IEEE International Conference on*, 2007, pp. 184–187.

[12] J.P. 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.

[13] C. Duxbury, J.P. 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.

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

[15] B. Thoshkahna and K.R. 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.

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

[17] L. Wan-Chi and C.-C.J. Kuo, "Improved linear prediction technique for musical onset detection," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2006, pp. 533–536.

[18] 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.

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

[20] 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.

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

[22] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE transactions on*, vol. 10, no. 5, pp. 293–302, 2002.

[23] E. Pavez and J.F. Silva, "Analysis and design of wavelet-packet cepstral coefficients for automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 814–835, 2012.

[24] E. Didiot, I. Illina, D. Fohr, and O. Mella, "A wavelet-based parameterization for speech/music discrimination," *Computer Speech and Language*, vol. 24, no. 2, pp. 341–357, 2010.

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

[26] A. Spanias, T. Painter, V. Atti, and J.V. Candy, "Audio Signal Processing and Coding," *Acoustical Society of America Journal*, vol. 122, pp. 15, 2007.

[27] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, pp. 386, 1958.

[28] J.L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[29] M.I. Jordan, "Artificial neural networks," pp. 112–127. IEEE Press, Piscataway, NJ, USA, 1990.

[30] K.J. Lang, A.H. Waibel, and G.E. Hinton, "A time-delay neural network architecture for isolated word recognition," *Neural networks*, vol. 3, no. 1, pp. 23–43, 1990.

[31] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, vol. 148, 2001.

[32] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[33] A. Graves, *Supervised sequence labelling with recurrent neural networks*, vol. 385, Springer, 2012.

[34] 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.

[35] "Ismir 2004 ballroom data set," 2004, http://mtg.upf.edu/ismir2004/contest/tempoContest/node5.html.

[36] 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.

[37] M.D. 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*, Lisbon, Portugal, 2007, ACM, pp. 623–632.