Proceedings of the Federated Conference on Computer Science and Information Systems pp. 147–152
DOI: 10.15439/2016F558
ACSIS, Vol. 8. ISSN 2300-5963

Analysis of time-frequency representations for musical onset detection with convolutional neural network

Bartłomiej Stasiak
Institute of Information Technology,
Lodz University of Technology,
ul. Wólczańska 215, 90-924
Poland
Email: bartlomiej.stasiak@p.lodz.pl

Jędrzej Mońko
Institute of Information Technology,
Lodz University of Technology,
ul. Wólczańska 215, 90-924
Poland
Email: render@widzew.net

***Abstract*—In this paper a convolutional neural network is applied to the problem of note onset detection in audio recordings. Two time-frequency representations are analysed, showing the superiority of standard spectrogram over enhanced autocorrelation (EAC) used as the input to the convolutional network. Experimental evaluation is based on a dataset containing 10,939 annotated onsets, with total duration of the audio recordings of over 45 min.**

I. Introduction

Onset detection is a well recognized and important problem in automatic music information retrieval. It directly addresses one of the most fundamental aspects of music – the time flow and novelty detection; abstracting from *what* and *how*, it concentrates on the *when* question and tries to answer it as precisely as possible. Interesting on its own, this problem is also fundamental in the analysis of many higher-level concepts, such as *rhytm*, *meter* or *tempo* [1]. In a broader context, sound attack analysis can also support other audio processing tasks, including i.a. audio to score alignment (score following), query-by-humming melody search, singing voice quality evaluation and speech analysis [2][3][4][5][6][7].

While trivial when looking at the musical score, note onset detection appears surprisingly complex when musical recordings with real instruments are considered, with all kinds of phenomena and effects like vibrato, glissando, varied dynamics, embouchure and articulation types, etc. As the result, the precise definition of the *onset time*, enabling to unambiguously locate it on the time axis may be difficult [8][1]. Various definitions, including Perceptual Onset Time (POT), Perceptual Attack Time (PAT), Acoustic Onset Time (AOT) and Note Onset Time (NOT) have been proposed [9][1] in order to highlight differences between the time when the onset is perceivable by a human listener, when it is measurable in the signal or when e.g. the *note-on* command is triggered by a MIDI synthesizer [10]. Presence of vibrato, glissandi or ornamentation, not to mention impulse noise or other distortions in low-quality recordings, may in fact render the problem ill-posed, which makes us resort to machine learning approaches for example-based definition of what an *onset* actually is.

This work was supported by the Faculty of Technical Physics, Information Technology and Applied Mathematics, Lodz University of Technology

II. Previous work

The classical approach to note onset detection is based on the *onset detection function* (ODF) constructed to detect novel events in the sound signal [8][11][12][13]. Typically, the signal waveform $x(t)$ is first split into a series of consecutive, usually overlapping time frames $x_n(t)$ with a windowing function applied to each of them:

$$x_n(t) = x(t)w(St - nh) \tag{1}$$

where $w(St - nh)$ is a windowing function stretched by a factor of frame size S and shifted by an integer, n-th multiple of the hop-size h between the consecutive frames. Discrete Fourier transform (DFT) is then computed, and the ODF construction may be based on either its magnitude spectrum [8], the phase spectrum [11] or both [12]. Obviously, the difference between the consecutive frames is considered, such as in the following simple example (ODF based on the *spectral flux* [1][14]):

$$ODF_{sf}(n) = \sum_k H(|X_k(n)| - |X_k(n-1)|) \tag{2}$$

where

$$H(x) = \frac{x + |x|}{2} \tag{3}$$

$$X_k(n) = \mathrm{DFT}(x_n(t))(k) \tag{4}$$

and where the half-wave rectifier function H is used to consider only positive differences, indicating new spectral components appearing in the signal. The onsets may be then easily detected by thresholding the ODF with a fixed threshold T or – more frequently – with a threshold based on moving mean or moving median.

978-83-60810-90-3/$25.00

It should be noted that some onsets (e.g. percussive ones) may be reliably detected also in the time domain by simply monitoring the signal energy. However, sound signal is generally better described in the frequency domain, as opposed to e.g. image processing, where the frequency domain methods have usually more limited and specialized applications [15][16]. For sound, spectral analysis if far more flexible and it opens possibilities of the construction of many specialized algorithms where the signal may be easily split into frequency bands, often distributed logarithmically according to human perception of the pitch. For example, Böck and Widmer proposed an onset detection algorithm with vibrato suppression called SuperFlux, where the input data is filtered with a bank of varying-length frequency domain triangular filters spaced equally in musical scale and where the maximum filter is applied to the resulting spectrograms in order to ignore minor pitch fluctuations [17]. It has been shown [18] that this approach enhances onset detection for bowed instruments playing both with and without the vibrato technique.

In contrary to classical onset detection methods, many recent works involve machine learning techniques – most notably the neural networks [19][20][21], although other data-driven techniques, such as Support Vector Machines (SVM) have also been applied [22]. The input data usually consists of a time-frequency representation of the sound signal, mapped non-linearly in the frequency domain according to a perceptual model. Böck *et al* [21] used a bank of triangular filters positioned at critical bands of the Bark scale to filter the STFT magnitude spectra, computed with three different window lengths in parallel. In this way the redundancy resulting from unnecessarily high frequency resolution of the STFT in the upper frequency range may be avoided. Hertz to Mel scale mapping [23] and constant-Q transform [20] have also been applied for similar reasons.

Several approaches have been proposed in which the *fusion* of many onset detection functions is applied. This is accomplished either on the feature-level by a set of pre-defined rules or a linear combination of ODFs [24], or in the form of the *score-level fusion* in which the decisions are taken on the basis of the already computed onsets [25][24]. Quintela *et al* [25] apply i.a. KNN- and SVM-based classifiers to the lists of pre-computed onset candidates and their locations in time. Recently, Stasiak *et al* [10] proposed to simultaneously use several ODF functions as the input to a multilayer perceptron with one output, playing *de facto* the role of a new "integrated" onset detector. In this way the neural network learns to merge the onset-related information from various sources, while not being forced to extract it explicitly from raw spectral data.

On the other hand, the recent progress in theory and practical applications of deep neural architectures enabled to successfully use the solutions developed by the image processing community also to directly process audio spectrograms, transforming the onset detection task into a problem similar to that of texture recognition. Apart from bidirectional long short-term memory neural networks (LSTM) [23] and recurrent neural networks (RNN) [21], the convolutional neural networks (CNNs) [26][27] proved to be especially useful here.

In this work we adopted the approach proposed in [27] to test the effectiveness of a convolutional network in the onset detection task using two different signal representations, namely the logarithmically scaled spectrogram and enhanced autocorrelation (EAC).

III. The proposed approach

A. Neural network architecture

The input to our network is a spectrogram fragment in the form of an image with 15 columns, representing 15 consecutive time frames and 80 rows, corresponding to 80 logarithmically distributed frequency bands (up to 16kHz). The initial audio files are sampled 44100Hz and the spectrogram parameters are: window size $N = 2048$, hop-size $K = 512$ samples, which yields time resolution of ca. 11.6 ms. The target is composed of a single value, indicating the distance of the onset from the middle frame of the current input image, similarly as in [10] (Fig. 1). If more onsets are present within the fragment, only the closest one is considered. In this way the network has to solve *regression* problem instead of binary classification (onset absent/present in the middle of the fragment). Preliminary experiments showed that it enhances the results significantly.

The network structure is as follows:

- Convolutional layer with ten rectangular filters of size: w×h = 7×3 with ReLU (Rectified Linear Unit) activation function and stride value of 1 in both directions (full overlap). Note that for input size of w×h=15×80 it yields w×h=9×78 output.
- Max-pooling layer with non-overlapping kernels of size: w×h = 1×3 (output size w×h = 9×26).
- Convolutional layer with twenty square filters of size w×h = 3×3 and stride value of 1 in both directions (output size w×h = 7×24).
- Max-pooling layer with non-overlapping kernels of size: w×h = 1×3 (output size w×h = 7×8).
- Inner product (i.e. fully connected) layer with 256 hidden neurons and ReLU activation function.
- Inner product (i.e. fully connected) layer with one output neuron and tanh (hyperbolic tangent) activation function.

The neural architecture basically follows the scheme proposed by Schlüter and Böck in [27] with some modifications concerning – apart from the aforementioned regression, replacing classification – mostly the type of nonlinearity of the layers. We agree with [27] that the rectified linear units in the first convolutional layer may play the role of the half-wave rectifier H function (cf. Eq. 3) helping to detect onset-related energy increases. Additionally, we use the same nonlinearity type for the fully connected layer, instead of sigmoidal units which proved to positively influence the learning process in our tests. We also change the unipolar sigmoid into tanh function in the output neuron, which leads to increasing the output range from $[0, 1]$ into $[-1, 1]$ (cf. Fig. 1, the top plot).

The last change influences the threshold which is applied to the output of the network in order to find the onset positions.

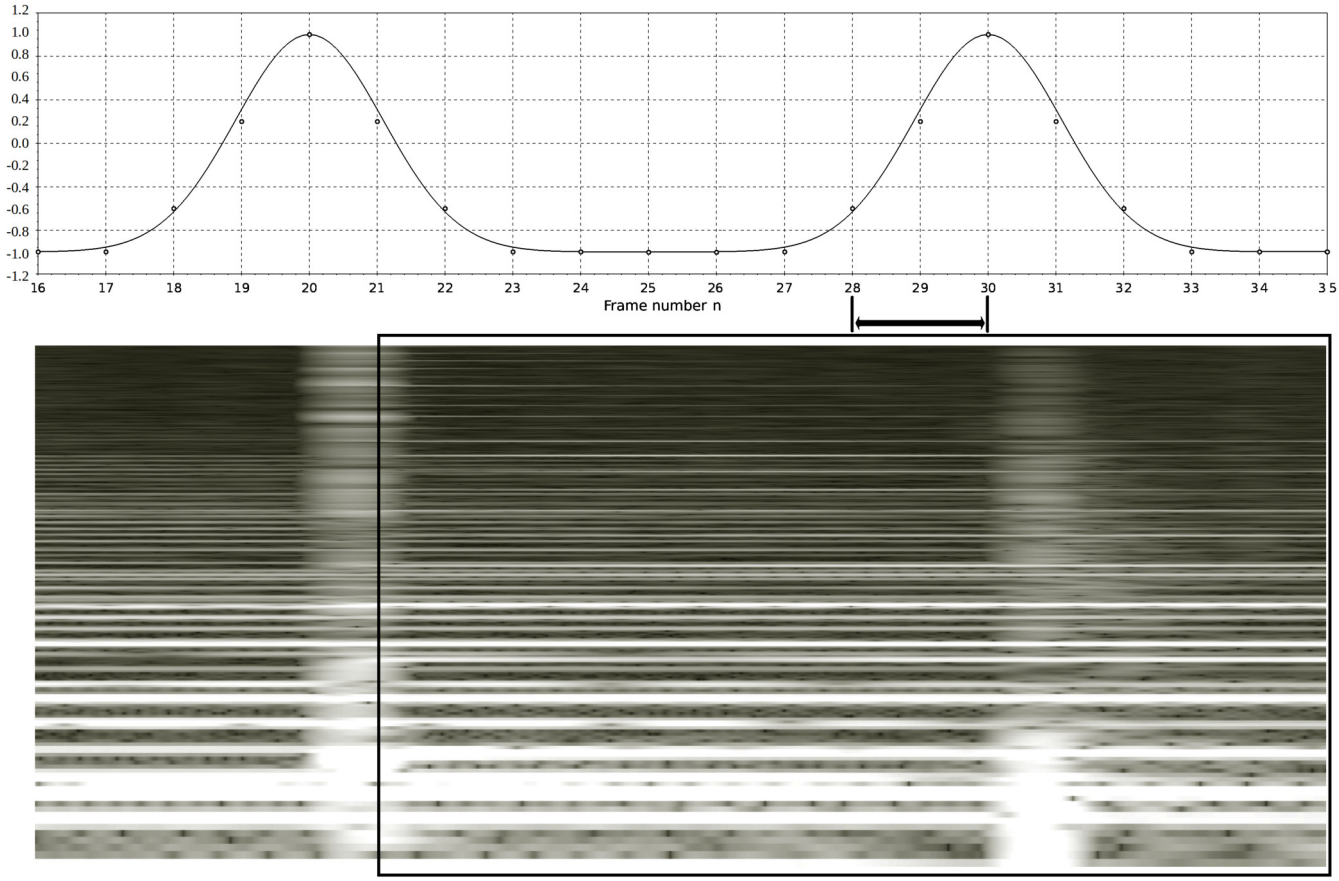

Fig. 1. *Spectrogram fragment enlarged (in a black box, lower plot) and the associated target values (top plot). The middle of this fragment (frame 28) is two frames apart from the onset (frame 30), so the target value for this fragment is -0.6 (top plot). Note, that the actual resolution of the image representing this fragment, that is fed to the network input is much lower (w×h=15×80 pixels)*

It should be mentioned here, that the output of the trained network may be treated as a classical ODF, with the difference, that a fixed threshold T may be used instead of moving mean or moving median, due to general lack of dependence on the signal energy. For the tanh activation function the optimal threshold value T_{opt} determined in our tests, i.e. the value maximizing the F-measure [27][10] was always lower than zero. After the thresholding, the peak-picking procedure is applied and peaks found within the range of 50ms relative to the actual onsets are treated as the properly detected ones.

Having denoted the correctly located onsets by **TP** (true positives), the assessment of the quality of the onset detection may be expressed in terms of *precision*, defined as the ratio: **TP/(TP+FP)**, and *recall*, defined as: **TP/(TP+FN)**. In our experiments we use the harmonic mean of precision and recall, known as the *F-measure*, as a "balanced" result of the onset detection procedure [10].

B. Audio material and data preparation

The dataset used in our experiments is a collection merged from several sources, including [8][28][29][20][30]. The total duration of all the audio files in our collection is over 45 min. and it contains 10,939 annotated onsets. The dataset has been divided at random into the train, test and validation subsets, containing 6236, 2520 and 2183 onsets, respectively. Complete files are assigned to either of the subsets (they are not split between the subsets).

For training, the spectrograms are cut into overlapping fragments which are then selected so that the obtained set is balanced, i.e. the number of "onset fragments" (for which the target value is non-zero) is equal to the number of "non-onset fragments". For testing, all possible fragments are used in an ordered sequence.

The main time-frequency representation (TFR) used in the experiments is the spectrogram, computed as explained in the previous section. In a separate test we use also enhanced autocorrelation (EAC) correlogram, calculated frame-by-frame in a similar way. EAC is an intermediate representation for the task of pitch estimation, thus also suitable for supporting onset detection with a degree of additional information on the input audio features related to melodic content. The procedure itself had been developed by Tolonen and Karjalainen [31] and (as the name implies) it is an extension of standard autocorrelation method. In our research we use EAC implementation operating in frequency domain for each frame of input signal using the

following processing scheme:

1) Transform a frame to frequency domain with Fourier transform. In this step, we use the same parameters as for the spectrogram computation – frame size of 2048 samples and frame step size (hop-size) of 512 samples.
2) Compute signal power
3) Take cube root of the resulting transform to compress magnitude in a non-linear manner. For normal autocorrelation the spectral coefficients are raised to the power of 2, however using the factor of 1/3 (cube root) of "generalised autocorrelation" is more suitable for the task of periodicity detection.
4) Clip all values below zero
5) Create a stretched copy (by factor of 2) of the values derived, and subtract it from the original (at step 4)
6) Clip all values below zero
7) Transform back to time domain with inverse Fourier transform

Steps 4-6 are performed for the purpose of peak pruning to improve pitch representation clarity, following Tolonen and Karjalainen's method. The procedure is applied to the signal frame-wise, yielding a correlogram, which can be processed further in a similar way as an ordinary spectrogram (Fig. 2).

C. Experimental evaluation

Caffe framework [32] has been used for training and testing the convolutional neural network presented in Sect. III-A. Separate validation set was used to determine the optimal model and to avoid overfitting. Stochastic gradient descent with momentum was used as the optimization strategy with mini-batch size of 1000 input spectrogram fragments. We used fixed momentum parameter of 0.4 and variable step size.

In the initial experiments we tested the influence of the activation function type on the results, as discussed in Sect. III-A. The results are presented in Table I.

TABLE I
THE RESULTS OF THE ONSET DETECTION TESTS

Experiment	F-measure
Original architecture based on [27]	82.13%
Our version with ReLU in the hidden layer and tanh in the output layer	83.35%
Our version trained on EAC correlograms instead of the spectrograms	73.10%

Although our modification enhanced the result by over one percent point, yet the EAC correlogram appeared definitely inferior to the spectrogram-based TFR. In the search of the potential reasons we conducted a series of additional tests in which we compensated for the potentially different annotation procedures in our heterogeneous dataset, by artificially shifting the onset positions by several multiples of the hop-size (from $-4 \times K$ to $4 \times K$). All the onsets in a given file were naturally shifted by the same displacement, but the displacement for each file was determined independently. Due to the latter fact, the figures presented in Table II obviously cannot be treated as the final objective results achieved by our network – they are rather indicators of its theoretical capabilities if some strict, uniform rules were applied for annotating the input files. They may also be used for a comparison of the spectrogram- and correlogram-based representations which again shows definite superiority of the first one. Figure 3 demonstrates the F-measure changes for varying values of the threshold T for the spectrogram-based input.

TABLE II
THE RESULTS OF THE ONSET DETECTION TESTS WITH ONSET SHIFTING

Experiment	F-measure
Our version with ReLU in the hidden layer and tanh in the output layer	88.62%
Our version trained on EAC correlograms instead of the spectrograms	81.13%

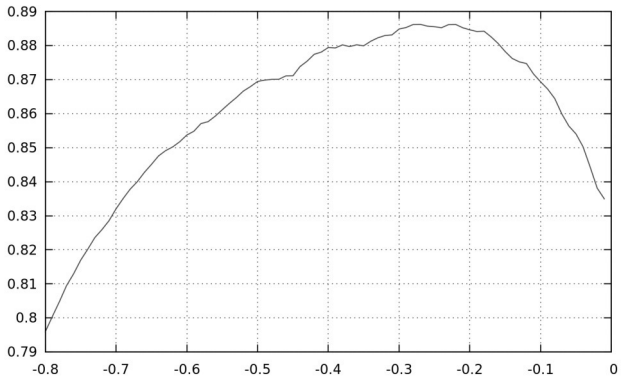

Fig. 3. *F-measure for varying values of the threshold T*

IV. DISCUSSION AND FUTURE WORKS

In this paper a convolutional neural network has been applied in the note onset detection problem. The obtained results demonstrate the superiority of the standard, spectrogram-based representation of the audio signal over the EAC correlogram. This observation confirms the potential of convolutional neural networks which are able to successfully extract useful information from the lower-level audio representation (spectrogram). The EAC correlogram, on the other hand, may be seen as a result of some more sophisticated processing, yielding more directly interpretable information related to pitch and melody content. However, this processing, although potentially useful from the human perspective, inevitably removes some information, which in consequence limits the potential of the convolutional neural network, eventually impairing the results.

The obtained results for the spectrogram-based input are satisfactory in terms of absolute onset detection rate. Enhancements might be searched for in increasing the precision of onset location (the annotations in the database used in the experiments should be manually checked and corrected to obtain more consistent annotation style [14]). Also, combining several time-frequency representations in a single spectrogram fragment, possibly computed with varied window size

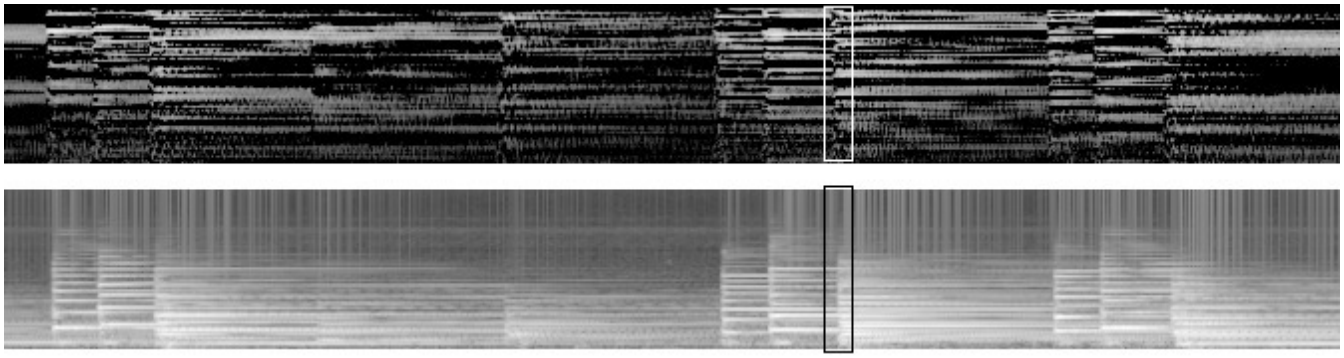

Fig. 2. *EAC correlogram (top) and spectrogram (bottom) of the same input file, with a bounding box around a fragment with an onset in the middle*

as proposed in [27], would probably lead to some further improvements.

ACKNOWLEDGMENT

We are truly grateful to Juan Pablo Bello, Sebastian Böck and Andre Holzapfel for making the annotated audio datasets available for our experiments.

REFERENCES

[1] A. Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley-IEEE Press, 2012.

[2] B. Thoshkahna and K. R. Ramakrishnan, "An onset detection algorithm for query by humming (QBH) applications using psychoacoustic knowledge." in *Proc. of 17th European Signal Processing Conference, EUSIPCO 2009*. IEEE, 2009, pp. 939 – 942.

[3] B. Stasiak, "Query by Singing/Humming (MIREX 2015). The Tune Follower," 2015. [Online]. Available: http://www.music-ir.org/mirex/abstracts/2015/BS2.pdf

[4] M. Purgina, A. Kuznetsov, and E. Pyshkin, "An approach for developing a mobile accessed music search integration platform," in *Proc. of Federated Conference on Computer Science and Information Systems, FedCSIS 2013*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds. IEEE, 2013, pp. 267–273.

[5] E. Półrolniczak and M. Kramarczyk, "Analysis of the sound attack in context of computer evaluation of the singing voice quality," in *Proc. of Federated Conference on Computer Science and Information Systems, FedCSIS 2015*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 5. IEEE, 2015. doi: 10.15439/2015F240 pp. 889–894.

[6] B. Stasiak and K. Rychlicki-Kicior, "Fundamental frequency extraction in speech emotion recognition," *Communications in Computer and Information Science*, vol. 287, pp. 292 – 303, 2012. doi: 10.1007/978-3-642-30721-8-29

[7] H. Wang and L. Wang, "Onset detection algorithm in voice activity detection for Mandarin," in *Proc. of Int. Conf. on Computer Science and Network Technology (ICCSNT)*. IEEE, 2013. doi: 10.1109/ICCSNT.2013.6967305 pp. 1148 – 1151.

[8] J. Bello, L. Daudet, S. Abdullah, C. Duxbury, M. Davies, and M. Sandler, "A Tutorial on Onset Detection in Music Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, September 2005. doi: 10.1109/TSA.2005.851998

[9] B. H. Repp, "Patterns of note onset asynchronies in expressive piano performance," *Journal of the Acoustical Society of America (JASA)*, vol. 100, no. 6, pp. 3917–3932, 1996. doi: 10.1121/1.417245

[10] B. Stasiak, J. Mońko, and A. Niewiadomski, "Note onset detection in musical signals via neural-network-based multi-ODF fusion," *International Journal of Applied Mathematics and Computer Science*, vol. 26, no. 1, pp. 203 – 213, 2016. doi: 10.1515/amcs-2016-0014

[11] P. Bello and M. Sandler, "Phase-based note onset detection for music signals," in *Proceedings of IEEE Conference on Acoustics, Speech, and Signal Processing ICASSP*, vol. 5, 2003. doi: 10.1109/ICASSP.2003.1200001 pp. 441–444.

[12] C. Duxbury, J. Bello, M. Davies, and M. Sandler, "Complex Domain Onset Detection For Musical Signals," in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, September 2003.

[13] J. Laroche, "Efficient Tempo and Beat Tracking in Audio Recordings," *Journal of the Audio Engineering Society (JAES)*, vol. 51, no. 4, pp. 226–233, 2003.

[14] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods." in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), 2012.*, 2012.

[15] V. Korzhik, G. Morales-Luna, A. Kochkarev, and I. Shevchuk, "Fingerprinting system for still images based on the use of a holographic transform domain," in *Proc. of Federated Conference on Computer Science and Information Systems, FedCSIS 2013*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds. IEEE, 2013, pp. 585–590.

[16] B. Stasiak and M. Yatsymirskyy, *Frequency Domain Methods for Content-Based Image Retrieval in Multimedia Databases*. Springer Berlin Heidelberg, 2009, pp. 137 – 166. ISBN 978-3-642-02196-1. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-02196-1_6

[17] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13)*, Maynooth, Ireland, September 2013, pp. 55–61.

[18] B. Stasiak and J. Mońko, "Analysis of Onset Detection with a Maximum Filter in Recordings of Bowed Instruments," in *Proceedings of the 138th Audio Engineering Society Convention*, May 2015. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=17695

[19] M. Marolt, A. Kavcic, and M. Privosnik, "Neural networks for note onset detection in piano music," in *Proceedings of the International Computer Music Conference*, 2002.

[20] A. Lacoste and D. Eck, "A Supervised Classification Algorithm for Note Onset Detection." *EURASIP Journal of Advanced Signal Processing*, pp. 153–153, 2007. doi: 10.1155/2007/43745

[21] S. Böck, A. Arzt, F. Krebs, and M. Schedl, "Online Real-time Onset Detection with Recurrent Neural Networks," in *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx 2012)*, September 2012.

[22] M. Davy and S. J. Godsill, "Detection of abrupt spectral changes using support vector machines. An application to audio signal segmentation." in *ICASSP*. IEEE, 2002. doi: 10.1109/ICASSP.2002.1005992 pp. 1313–1316.

[23] F. Eyben, S. Böck, B. Schuller, and A. Graves, "Universal Onset Detection with Bidirectional Long Short-Term Memory," in *Neural Networks, 11 th International Society for Music Information Retrieval Conference (ISMIR 2010)*, 2010, pp. 589–594.

[24] M. Tian, G. Fazekas, D. A. A. Black, and M. Sandler, "Design and Evaluation of Onset Detectors Using Different Fusion Policies," in *15th International Society of Music Information Retrieval (ISMIR) Conference*, 2014, pp. 631–636.

[25] N. D. Quintela, A. P. Giménez, and S. T. Guijarro, "A Comparison of Score-level Fusion Rules for Onset Detection in Music Signals." in *Proceedings of 10th International Society for Music Information Retrieval Conference ISMIR09*, October 2009, pp. 117–121.

[26] J. Schlüter and S. Böck, "Musical Onset Detection with Convolutional Neural Networks," in *6th International Workshop on Machine Learning and Music (MML)* , 2013.

[27] J. Schlüter and S. Bock, "Improved Musical Onset Detection with Convolutional Neural Networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy, 2014. doi: 10.1109/ICASSP.2014.6854953

[28] L. Daudet, G. Richard, and P. Leveau, "Methodology and Tools for the evaluation of automatic onset detection algorithms in music." in *ISMIR*, 2004, pp. 72–75.

[29] A. Holzapfel, Y. Stylianou, A. C. Gedik, and B. Bozkurt, "Three dimensions of pitched instrument onset detection." *IEEE Trans. Audio, Speech & Language Processing*, vol. 18, no. 6, pp. 1517–1527, 2010. doi: 10.1109/TASL.2009.2036298

[30] J. Glover, V. Lazzarini, and J. Timoney, "Real-time detection of musical onsets with linear prediction and sinusoidal modeling." *EURASIP J. Adv. Sig. Proc.*, vol. 2011, p. 68, 2011. doi: 10.1186/1687-6180-2011-68

[31] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, Nov 2000. doi: 10.1109/89.876309

[32] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.