DE GRUYTER OPEN

Int. J. Appl. Math. Comput. Sci., 2016, Vol. 26, No. 1, 203–213
DOI: 10.1515/amcs-2016-0014

amcs

NOTE ONSET DETECTION IN MUSICAL SIGNALS VIA NEURAL–NETWORK–BASED MULTI–ODF FUSION

BARTŁOMIEJ STASIAK [a,*], JĘDRZEJ MOŃKO [a], ADAM NIEWIADOMSKI [a]

[a] Institute of Information Technology
Łódź University of Technology, ul. Wólczańska 215, 90-924 Łódź, Poland
e-mail: bartlomiej.stasiak@p.lodz.pl

The problem of note onset detection in musical signals is considered. The proposed solution is based on known approaches in which an *onset detection function* is defined on the basis of spectral characteristics of audio data. In our approach, several onset detection functions are used simultaneously to form an input vector for a multi-layer non-linear perceptron, which learns to detect onsets in the training data. This is in contrast to standard methods based on thresholding the onset detection functions with a moving average or a moving median. Our approach is also different from most of the current machine-learning-based solutions in that we explicitly use the onset detection functions as an intermediate representation, which may therefore be easily replaced with a different one, e.g., to match the characteristics of a particular audio data source. The results obtained for a database containing annotated onsets for 17 different instruments and ensembles are compared with state-of-the-art solutions.

Keywords: note onset detection, onset detection function, multi-layer perceptron, multi-ODF fusion, NN-based fusion.

1. Introduction

Segmentation may be deemed one of the most important skills attributed to intelligence, either human or artificial. Assigning significance to some spatially or temporally correlated groups of data in an image or a sound file is an elementary step of analysis, providing grounds for feature extraction, description and, eventually, comprehension.

A fundamental stage in audio segmentation process is *onset detection* or, especially in music information retrieval (MIR), *note onset detection*. It is used as a starting point in numerous practical applications, including rhythm and tempo analysis (Laroche, 2003; Peeters, 2005), query-by-humming (QbH) music search engines (Huang *et al*., 2008; Typke *et al*., 2007), support systems in music education (Zhang and Wang, 2009; Yin *et al*., 2005) and parametric audio coding (Bartkowiak and Januszkiewicz, 2012).

Note onsets are tightly related to attack transients in musical signals. This is due to the fact that the sound produced by a musical instrument is a non-stationary signal in a short period of time after excitation occurs. Detection of transients, in particular attack transients, is applicable in musical signal processing due to the fact that during transient states the magnitude and the phase of a signal tend to change rapidly. However, the precise definition of the *onset time*, making it possible to unambiguously locate it on the time axis, is not a straightforward task (Bello *et al*., 2005; Lerch, 2012). Various definitions, including perceptual onset time (POT), perceptual attack time (PAT), acoustic onset time (AOT) and note onset time (NOT), have been proposed (Repp, 1996; Lerch, 2012) in order to highlight differences between the time when the onset is perceivable by a human listener, when it is measurable by audio monitoring devices, and simply when the *note-on* command is triggered by a MIDI synthesizer.

Analysis of polyphonic music is especially difficult, due to natural limitations of performers' precision in playing several notes simultaneously. Moreover, the onset-specific type of change in the temporal and spectral characteristics of a sound varies significantly for different instruments and types of articulation. For example, pitched non-percussive (PNP) sounds, as those produced by bowed instruments, are generally considered more difficult to analyze than pitched/non-pitched percussive ones (PP/NPP), as intensity-related features may be not

*Corresponding author

sufficient for successful segmentation (Zhang and Wang, 2009; Collins, 2005). Finally, it should be remembered that musical instruments are basically modelled by complex systems of linear and non-linear ordinary as well as partial differential equations with time-varying parameters (Rabenstein and Petrausch, 2008), and the complete description of physical phenomena occurring in onset-related transient states is definitely nontrivial. All these factors make the task of building a universal onset detector a real challenge, which justifies searching for new, machine-learning-based methods which would be able to deal with the uncertainties inherent in formulation of the problem.

The classical approach in the onset detection task is composed of construction of an *onset detection function* (ODF), also known as a novelty function, and picking the peaks of the ODF, which indicate the occurrence of something new in the signal (Bello *et al*., 2005; Bello and Sandler, 2003; Duxbury *et al*., 2003; Laroche, 2003). In this work we propose a novel solution, combining the ODF-based approach with machine learning. One of the key advantages of our method (*NN-based multi-ODF fusion*) is simultaneous application of many ODFs, allowing covering a broad range of onset-relevant information.

The remainder of the paper is organized as follows. The next section presents methods and algorithms proposed in the literature, with an emphasis on neural-network-based solutions (Section 2.3). In this context, our approach is presented (Section 3), along with the description of the audio dataset selected for testing, details of the data preparation procedure and testing schemes. The obtained results are displayed and discussed (Section 4), and some conclusions and future perspectives are formulated in the last section.

2. Previous work on onset detection

2.1. Methods based on onset detection functions. Most of the ODF construction methods found in the literature utilize information about the magnitude and/or phase of STFT (short-time Fourier transform) frequency bins in consecutive frames, for finding spectrum changes indicating note onset occurrences. For example, one of the simplest approaches, known as the *spectral flux*, is based on a sum of half-wave rectified differences between the k-th magnitude spectrum bins of two consecutive STFT frames (Bello *et al*., 2005):

$$\mathrm{SF}(n) = \sum_k \mathrm{hwr}(|X_k(n)| - |X_k(n-1)|) \,, \quad (1)$$

where $X_k(n)$ is the k-th complex frequency bin of the n-th frame and

$$\mathrm{hwr}(x) = \frac{x + |x|}{2}$$

is the *half-wave rectifier* function, so that only positive changes in the magnitude are taken into account.

Most of the methods which involve information on phase changes rely on the differences between the predicted and actual phases of each frequency bin. This can be defined as

$$d\varphi_k(n) = \mathrm{princarg}[\varphi_k(n) - 2\varphi_k(n-1) + \varphi_k(n-2)] \,,$$

where $\varphi_k(n)$ is the k-th frequency bin of the n-th STFT frame and the princarg operator maps the argument to the $[-\pi, \pi]$ range. In the case of musical signals, the ODF depending solely on phase information may be sensitive to changes in all spectrum bins regardless of their magnitude. Therefore, it is worth combining the information on magnitude and the phase, e.g., by weighing phase deviation coefficients $d\varphi_k(n)$ by magnitude changes:

$$\mathrm{WPD}(n) = \sum_k |(|X_k(n)| - |X_k(n-1)|) \cdot d\varphi_k(n)| \,. \quad (2)$$

A more sophisticated method based on the phase spectrum was proposed by Bello and Sandler (2003), who used phase deviation coefficients to build a bin histogram of phase deviations for every STFT frame. Then the result may be calculated with some statistical characteristics (e.g., kurtosis) of such a distribution:

$$\mathrm{PHK}(n) = \mathrm{Kurt}(h(d\varphi(n))) \,, \quad (3)$$

where h is the bin histogram of the phase deviations.

There also exist methods which define the detection function on the basis of the complex spectrum, thereby taking into account both the amplitude and the phase. Referring to Duxbury *et al*. (2003), the detection function may be formulated as follows:

$$\mathrm{CD}(n) = \sum_k |\hat{S}_k(n) - S_k(n)| \,, \quad (4)$$

where $|\hat{S}_k(n) - S_k(n)|$ is the magnitude of the complex difference between the expected (predicted) and the actual k-th frequency bin of n-th STFT frame, where

$$\hat{S}_k(n) = |S_k(n-1)| e^{j(2\varphi_k(n-1) - \varphi_k(n-2))} \,.$$

The resulting detection functions are processed with adaptive thresholding and peak-picking algorithms. A moving average or a moving median is usually preferred over a fixed threshold as it can follow the dynamics of a sound (Duxbury *et al*., 2003; Böck *et al*., 2012). Additionally, some methods for controlling the salience of a peak are often applied (Dixon, 2006). Nevertheless, unequivocal determination of the onsets is far from trivial, and both false positives (FP: onsets reported in places where no onset actually appears in the recording) and false negatives (FN: actual onsets that have not been reported)

are inherent to practically all the approaches proposed so far.

Having denoted the correctly located onsets by TP (true positives), the assessment of 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). Note that too low a threshold value leads to reporting most of peaks, including the irrelevant ones, and thus it results in excellent recall (low FN) but poor precision (high FP). The opposite outcome (low recall and high precision) is expected for too high a threshold value, overshooting many relevant peaks. The harmonic mean of precision and recall, known as the *F-measure*, is therefore often reported as a "balanced" result of the onset detection procedure (Dixon, 2006; Böck *et al*., 2012).

2.2. Multi-ODF fusion. A separate research direction, especially related to our approach, is *fusion* of several onset detection functions. This is accomplished either on the feature-level by a set of pre-defined rules or a linear combination of ODFs (Tian *et al*., 2014), or in the form of the *score-level fusion* in which the decisions are taken on the basis of the already computed onsets (Quintela *et al*., 2009; Tian *et al*., 2014). However, despite the apparent similarities to our solution (cf. Section 3), the differences are indeed very significant. Tian *et al*. (2014) deliberately refrain from using machine learning, while Quintela *et al*. (2009) although they apply, e.g., KNN- and SVM-based classifiers, operate on a completely different representation of input data in the form of lists of pre-computed onset candidates and their locations in time. In this context, our neural-network-based approach relying on unprocessed raw ODF values presents an alternative point of view on the onset detection problem.

2.3. Methods based on machine learning. The popularity of machine learning applications for onset detection is growing rapidly with some excellent results reported in recent research. Neural networks are the tool of choice (Lacoste and Eck, 2007; Böck *et al*., 2012), although other data-driven techniques have also been used (Davy and Godsill, 2002). The input data usually consist of a time-frequency representation of the sound signal, mapped non-linearly in the frequency domain according to a perceptual model. Böck *et al.* (2012) 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 (Eyben *et al*., 2010) and the constant-Q transform (Lacoste and Eck, 2007) have also been applied for similar reasons. Nevertheless, the problem of dimensionality reduction of the data used as the neural network input is not fully resolved, yet forcing system designers to apply special preprocessing methods, including, e.g., random sampling of the input window along time and frequency axes (Lacoste and Eck, 2007).

The structure of the neural networks used in the onset detection problem has often been subjected to extensive research, and some non-standard approaches have also been proposed. For instance, a multi-net approach proposed by Lacoste and Eck (2007) is based on merging the results obtained from several networks, each trained with a different set of hyper-parameters, by means of an additional "output" neural network followed by a peak-picking procedure. Apart from the standard questions regarding the number of hidden layers and hidden neurons, several different NN types, including the recurrent neural network (RNN), the feed-forward convolutional neural network (CNN) and the LSTM (long short-term memory) neural network, have been considered (Böck *et al*., 2012; Eyben *et al*., 2010; Schlüter and Böck, 2014).

3. Our solution: NN-based multi-ODF fusion

In our approach a neural network is applied in a different way for solving the onset detection problem. Instead of taking a pre-processed spectrogram as the raw input, we compute several onset detection functions and put their values to the input of the neural network. The network summarizes the information from all the ODFs and generates its own onset probability estimation on this basis (Fig. 1).

This approach follows the standard division of a pattern recognition system into feature extraction and classification blocks. The main role of the feature extraction block is to compute the onset detection functions which basically employ much more problem-specific *a priori* knowledge compared with approaches in which this knowledge has to be learnt directly from spectral data. In this way the construction of the classifier itself may be simplified and the input space dimensionality may be kept reasonably low. This is especially important as multi-dimensional data need more training examples, which—in the case of the onset detection problem—implies a laborious process of manual annotation of audio files (Daudet *et al*., 2004).

3.1. Dataset. The dataset collected by Pierre Leveau (Daudet *et al*., 2004) was chosen to test the effectiveness of our solution. The collection contains 17 audio files representing a variety of music styles and instruments. It was annotated by three expert listeners for the total number of over 670 onsets, reported in the corresponding ground-truth files. It should be noted that

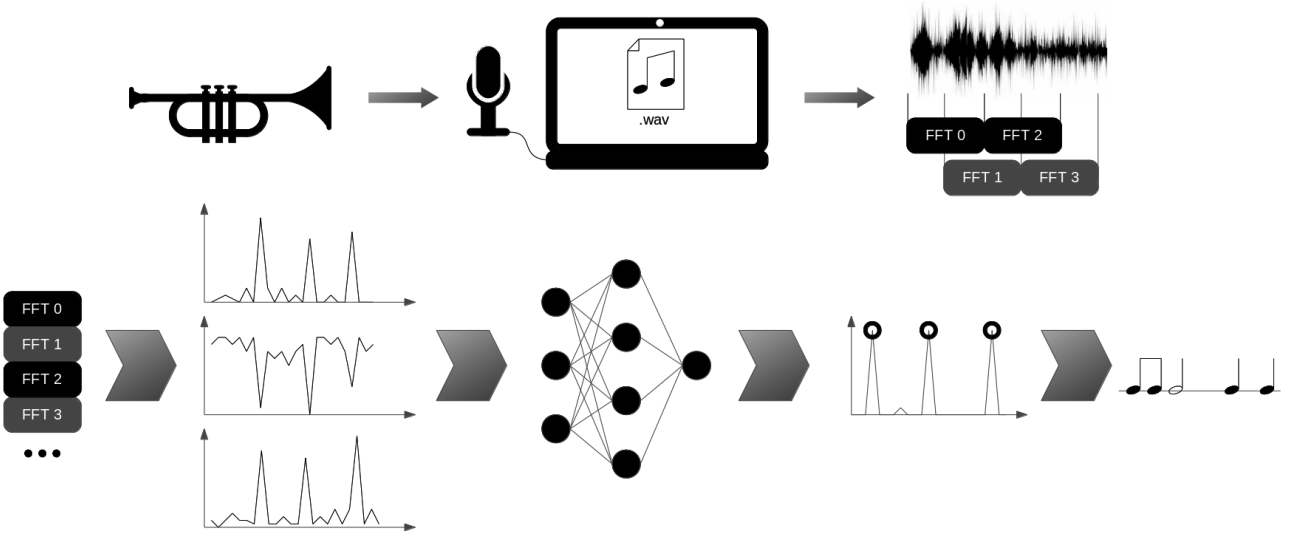


Fig. 1. Processing steps of our onset detection system. Top: audio data acquisition and spectrogram computation, bottom: construction of several onset detection functions, neural network training and thresholding the output.

all the three annotations had to be consistent for any given onset to be included in the database. For this reason, some of the onsets are missing from the ground-truth files if their timing differed between the annotators by more than a predefined value (100 ms).

3.2. Data preparation. A basic tool used in our solution is a multi-layer perceptron (MLP) with one hidden layer and a non-linear (unipolar sigmoid) activation function. As has been stated before, the input of the neural network is based on the data obtained from several onset detection functions aligned in time and sampled uniformly within a sliding window. In fact, no explicit sampling is necessary, as the ODFs are already extracted from the audio signal on the per-frame basis. In our approach the original audio files, recorded at $f_s = 44.1$ kHz, were cut into frames of size $N = 2048$ with a half-frame overlap, which resulted in computing the ODF samples every K ms, where

$$K = \frac{1}{f_s}\frac{N}{2} \approx 23.22\,. \tag{5}$$

Four onset detection functions defined with the formulas (1)–(4) were included into the study (Fig. 2, left column), although any other type and number of ODFs may be applied as well. The sliding window with a fixed number of $n_s = 5$ samples for each of the four ODFs is used. A single input vector is therefore composed of 20 samples plus additional four values computed as arithmetic means of $n_m = 21$ ODF samples in the neighborhood of the sliding window. In this way we incorporate the concept of a moving mean into our solution, so that the classifier may benefit also from the long-term information related to the average level of the signal in a given interval. Finally, the neural network has 24 inputs, and the input vector for a given location of the sliding window, denoted by n, takes the form of a concatenation of four vectors (cf. Fig. 2, right column):

$$\boldsymbol{x}(n) = [\boldsymbol{v}_{\mathrm{SF}}(n),\ \boldsymbol{v}_{\mathrm{WPD}}(n),\ \boldsymbol{v}_{\mathrm{PHK}}(n),\ \boldsymbol{v}_{\mathrm{CD}}(n)]\ , \tag{6}$$

defined as

$$\begin{aligned}
\boldsymbol{v}_{\mathrm{SF}}(n) =& \Big[\mathrm{SF}\,(n-2)\,,\ \mathrm{SF}\,(n-1)\,,\ \mathrm{SF}\,(n)\,,\ \ldots \\
& \mathrm{SF}\,(n+1)\,,\ \mathrm{SF}\,(n+2)\,,\ \overline{\mathrm{SF}}(n)\Big], \\
\boldsymbol{v}_{\mathrm{WPD}}(n) =& \Big[\mathrm{WPD}\,(n-2)\,,\ \mathrm{WPD}\,(n-1)\,, \\
& \mathrm{WPD}\,(n)\,,\ \ldots, \mathrm{WPD}\,(n+1)\,, \\
& \mathrm{WPD}\,(n+2)\,,\ \overline{\mathrm{WPD}}(n)\Big], \\
\boldsymbol{v}_{\mathrm{PHK}}(n) =& \Big[\mathrm{PHK}\,(n-2)\,,\ \mathrm{PHK}\,(n-1)\,, \\
& \mathrm{PHK}\,(n)\,,\ \ldots, \mathrm{PHK}\,(n+1)\,, \\
& \mathrm{PHK}\,(n+2)\,,\ \overline{\mathrm{PHK}}(n)\Big], \\
\boldsymbol{v}_{\mathrm{CD}}(n) =& \Big[\mathrm{CD}\,(n-2)\,,\ \mathrm{CD}\,(n-1)\,,\ \mathrm{CD}\,(n)\,, \\
& \ldots,\ \mathrm{CD}\,(n+1)\,,\ \mathrm{CD}\,(n+2)\,,\ \overline{\mathrm{CD}}(n)\Big],
\end{aligned} \tag{7}$$

where the last element of the vector $\boldsymbol{v}_{\mathrm{SF}}(n)$ is computed as

$$\overline{\mathrm{SF}}(n) = \sum_{k=n-\lfloor n_m/2 \rfloor}^{n+\lfloor n_m/2 \rfloor} \mathrm{SF}(k)\ , \tag{8}$$

and similarly for the remaining three vectors.

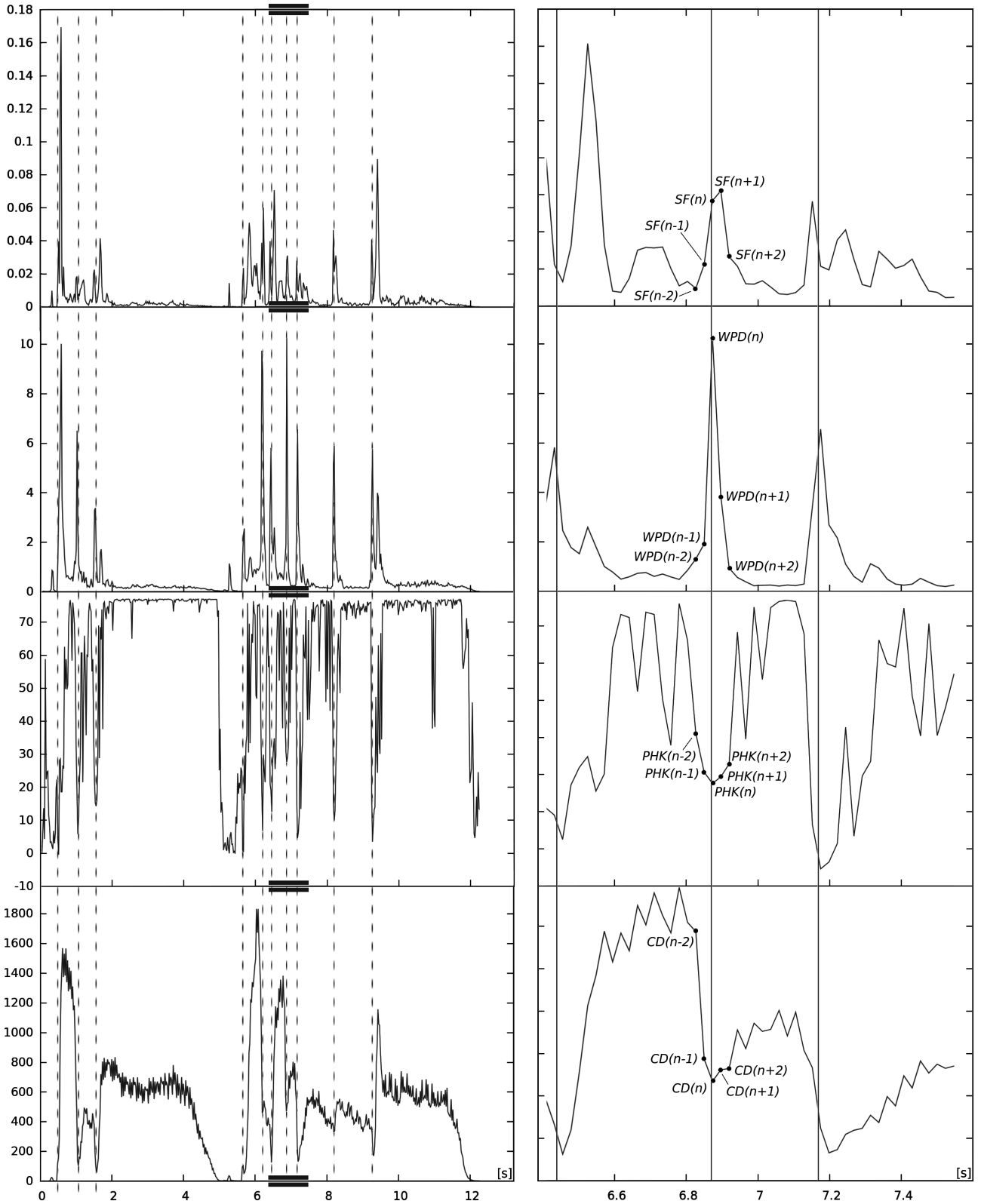


Fig. 2. Left column: four ODFs of a sax solo recording (G. Gershwin, *Summertime*, from *Porgy and Bess*—the beginning) and their selected fragment, marked with black rectangles in the left column, shown magnified in the right column. The vertical lines mark the ground-truth onsets.

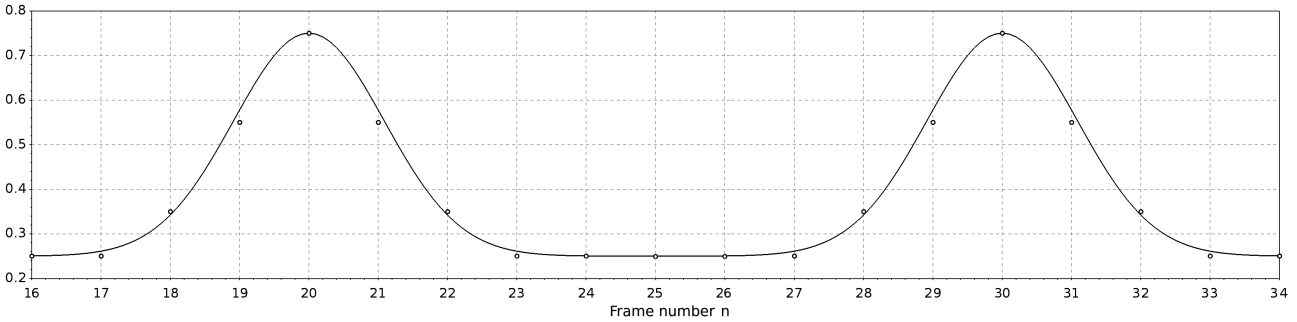


Fig. 3. Target values (the circles) for a fragment of an audio file with onsets appearing in the middle of the 20th and the 30th frame. The continuous line represents the ideal model (Gaussian curve) of the onsets.

At the output of the network, we expect a single value indicating the probability that an onset appears in the center of the sliding window at a given position, i.e., within the n-th frame, assuming the input vector in the form defined by Eqn. (6). However, a binary response (onset present/not present) may lead to misclassification if the onset in the ground-truth data appears one frame before or after the actual ODF peak, which may easily happen due to unavoidable imprecision of the music annotation process. We therefore decided to define a soft condition for the onset presence, in which the target output value of the network is modeled as a Gaussian curve centered at the n-th frame. After some simplifications and rounding, the consecutive target values for an onset appearing in the n-th frame are set to (Fig. 3)

$$\begin{aligned} t(n) &= 0.75, \\ t(n \pm 1) &= 0.55, \\ t(n \pm 2) &= 0.35, \\ t(n \pm 3) &= 0.25, \\ t(n \pm 4) &= 0.25, \\ &\vdots \end{aligned} \tag{9}$$

We decided to limit the range of output values to [0.25, 0.75] instead of [0, 1] because the unipolar sigmoid used as the neural activation function is unable to reach the endpoints of the second of these intervals, which might impede the learning process (Bishop, 1995).

The output of the neural network is treated as another onset detection function for which the peak-picking and thresholding procedures must be applied (cf. Section 2.1). The obvious advantage with respect to the original ODFs used to construct the input data is that there is no correspondence to the energy of the input signal, and hence a fixed threshold

$$T \in (0.25, 0.75) \tag{10}$$

may be used instead of the moving average or median. We also do not have to consider the characteristics of each individual ODF, such as whether the onsets are indicated by local maxima or minima.

3.3. Train/test procedures. Two training/testing schemes were applied:

1. In the first scheme, one instrument was removed from the dataset and all the remaining 16 were used to train the network (1-vs-all scheme). After the training had been finished, the removed instrument was used to test the network and only the results for this instrument ("unknown" in the learning phase) were reported. This procedure was repeated 17 times, so that each instrument was treated as the "unknown" one exactly once.

2. In the second scheme, a single audio file was used both for training and for testing in the 10-fold cross-validation procedure (c-v scheme). In this case all the remaining instruments were deliberately ignored and only a part (1/10) of the recording of the chosen one was treated as the "unknown" test data. The training was repeated 10 times, so that each 1/10 of the file was treated as the "unknown" one exactly once. The arithmetic means of all these ten folds were reported and the whole procedure was repeated for the remaining audio files.

The first scheme allows obtaining a universal onset detector, trained on a variety of sound sources. The task is more difficult here, as no data from the tested recording are used in the training stage and the obtained results for some instruments may be suboptimal if their characteristics differ much from the general "population". One particular problem that was encountered during the tests was that the results obtained for the clarinet and the saxophone were significantly delayed with respect to the annotated onsets in the ground-truth files (Fig. 4). This observation corresponds to the delay of the signal energy increase with respect to the beginning of embouchure, which is specific to some woodwind instruments.

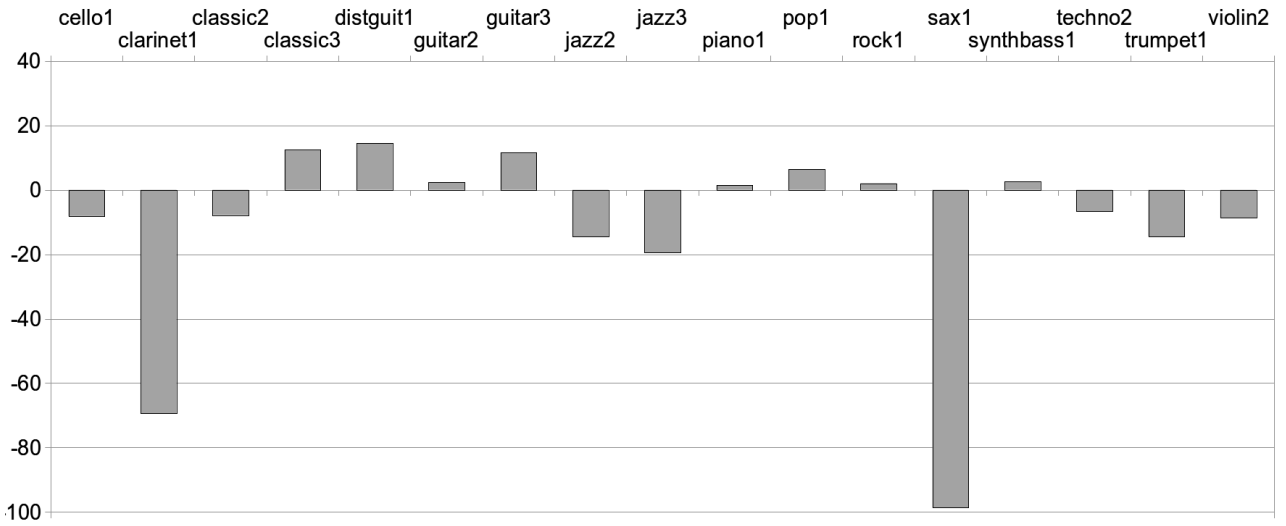


Fig. 4. Relative differences in milliseconds between the annotations in the ground-truth files and the results obtained in the first testing scheme (1-vs-all). For each instrument the optimal shift of all the onset positions, yielding the best F-measure, is reported.

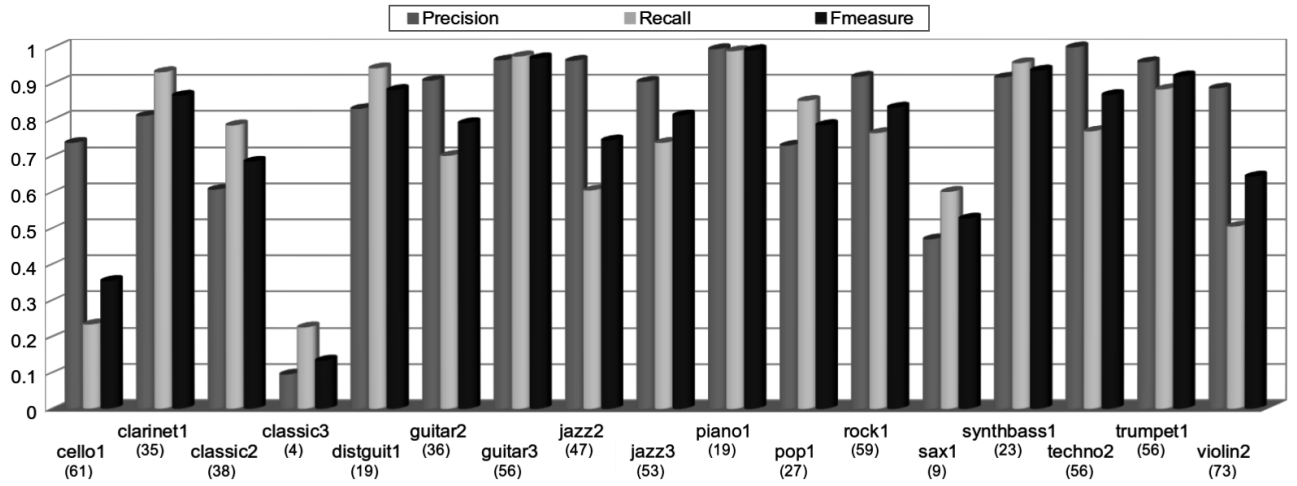


Fig. 5. Results: the first testing scheme (one instrument vs. all the others). The numbers in brackets represent the total number of annotated onsets for each instrument.

4. Results and discussion

The results obtained in the first testing scheme (1-vs-all), after pre-shifting the ground-truth onsets accordingly, are presented in Fig. 5. For each tested instrument, the training was repeated 10 times, and the average value is reported. These results were obtained after 100 training epochs in the off-line mode (Bishop, 1995). In fact, due to the large number of input vectors (ca. 10000), resulting from application of the sliding window (Fig. 2) to all the recordings from the training set, as few as 25 epochs were enough to reach the overall F-measure value of ca 78% (Table 1). Several networks with various numbers of hidden neurons between 15 and 80 were tested, but the variation in the outcomes was relatively low. The presented results were obtained for 30 hidden units.

The results obtained in the second scheme (c-v) are shown in Fig. 6. Here the data used for testing (1/10 of the recording of a given instrument) were also disjoint from the training data (the remaining 9/10). The train/test cycle for a single instrument was repeated 10 times until every fragment of the recording had been used exactly once for testing. The value reported is the arithmetic mean of all the ten folds. The number of training vectors is naturally much lower here compared with the first testing scheme: from 233 (*distguit1* file) to 1169 (*clarinet1*) per each fold and therefore the number of epochs must be appropriately greater. The fixed number of 5000 epochs was set for all the instruments, although the learning dynamics and the speed of convergence varied greatly in many cases. This

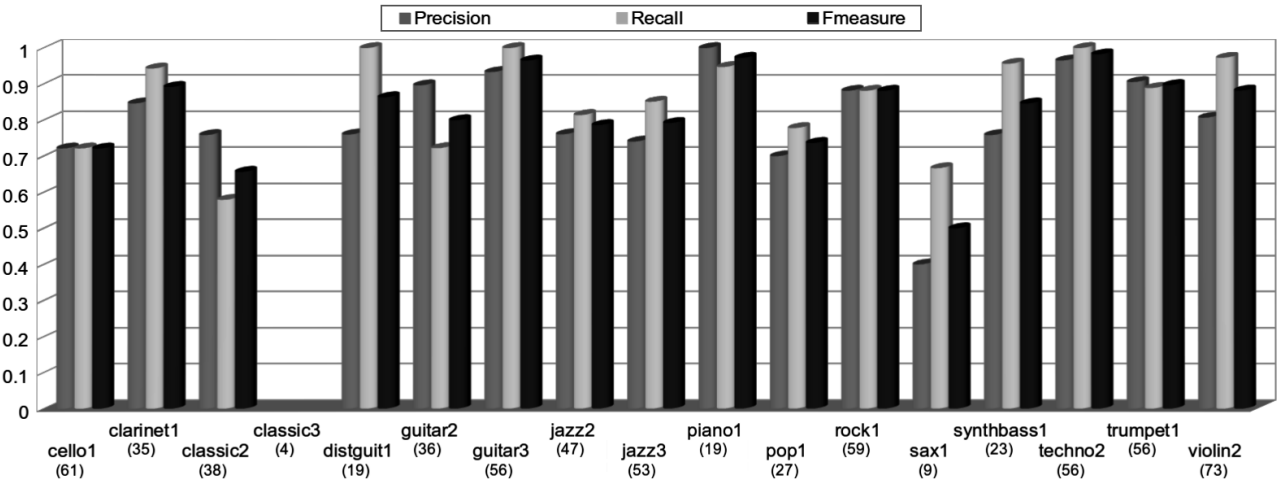


Fig. 6. Results: the second testing scheme (cross-validation: 1/10 of a recording vs. the remaining 9/10). The numbers in brackets represent the total number of annotated onsets for each instrument. Note that the number of onsets in the *classic3* dataset (four) appeared too small to successfully train the classifier for this case.

variability was observed during the tests in the obtained sequences of values of the training error E, defined as

$$E = \sqrt{\frac{1}{M}\sum_{n=1}^{M} |y(n) - t(n)|^2}, \tag{11}$$

where M is the number of available audio frames, $t(n)$ is the expected target value (Eqn. 9) and $y(n)$ is the actual value obtained at the output of the neural network for the n-th frame.

For example, the error value E after 5000 epochs for the piano and cello recordings reached the levels of 0.038 and 0.15, respectively, indicating the relative complexity of detection of the pitched, non-percussive cello onsets. Applying an individual approach to each instrument and using more flexible stopping criteria to control the generalization error (e.g., with a validation set) would supposedly lead to a further improvement of the results.

Independence of the output range of the network on the energy of the signal is an advantage, allowing usage of a single, fixed value T (Eqn. (10)) to threshold the output of the network. However, this value still has to be appropriately set in order to achieve the desired precision and recall levels. Maximization of the F-measure required the threshold value of 0.4 in the first testing scheme and 0.43 in the second one (Fig. 7). Comparison of the obtained values of precision, recall and the F-measure for both testing schemes is presented in Table 1.

4.1. Discussion. The general expectations on the superiority of the second testing scheme (c-v) are obviously met, as can be seen in Table 1. Training the network on the same type of data as those used for testing

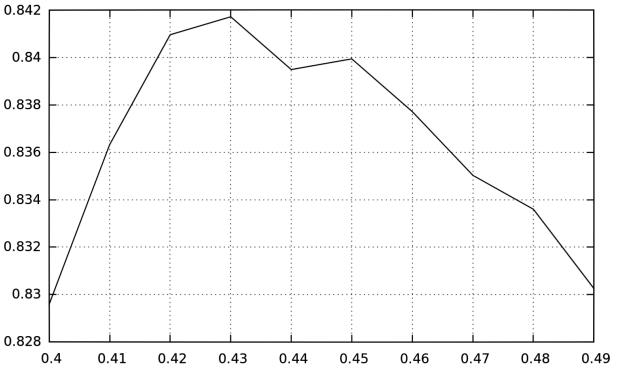


Fig. 7. Relation of the F-measure (ordinate) to the MLP output threshold (abscissa) for the second testing scheme (c-v).

leads to obtaining a more specialized and more effective onset detector. This is very well visible in the case of bowed instruments (cello and violin), which are of a very specific (PNP) onset type. Their F-measure was 0.354 and 0.643 in the first testing scheme and 0.721 and 0.882 in the second one. Low values of recall can be also observed in Fig. 5 for these two instruments, which may suggest that

Table 1. Overall results of the onset detection tests.

	Scheme 1 (1-vs-all)	Scheme 2 (c-v)
Correctly detected onsets	486	569
Precision	0.856	0.821
Recall	0.724	0.863
F-measure	0.785	0.842

Table 2. Results for ODF subgroups, sorted by the F-measure ('X' means that the corresponding ODF was included in the input set).

SF	WPD	PHK	CD	F-measure
X		X		0. 490
	X	X		0. 657
X	X	X		0. 658
X	X			0. 687
X	X		X	0. 729
X			X	0. 732
	X		X	0. 734
		X	X	0. 735
X		X	X	0. 774
	X	X	X	0. 781

the threshold of 0.4 used in the first scheme is too high in these cases. The PNP onsets are sufficiently different from most of the other onsets (used for training) to make the network "hesitate" and generate lower output values.

Comparing the results in Figs. 5 and 6, we can observe that for some instruments (e.g., synthbass) the onset detector trained on the other recordings performs better. The explanation may be that these instruments have relatively few onsets in the annotated ground-truth files, so the network simply does not have sufficient data for building a proper model of an onset when no other recordings are used (the second scheme). This is best seen in the case of the *classic3* file, containing only four annotated onsets, which results in zero values of both recall and precision. This file is specific also because it contains a fragment of orchestral music with very soft, slow onsets, presenting substantial problems to human annotators. Due to the imprecise timing, a significant number of the actual onsets were not included in the ground-truth data (cf. Section 3.1), leading to extremely low precision values also in the first testing scheme (Fig. 5). This may be, however, regarded more as a demonstration of the fundamental ambiguities underlying the formulation of the onset-detection problem in general.

The relative influence of the individual onset detection functions was evaluated in a separate group of tests in which the input vectors were reduced to contain the values of only two or three ODFs. The first testing scheme (one instrument vs. all the others) was applied for these tests. For each subgroup the threshold yielding the highest F-measure value was sought (the threshold values fell within the range 0.36–0.42). The results presented in Table 2 indicate that the complex-domain spectrum (Eqn. (4)) contains the most useful onset-related information. Its removal leads to a rather rapid drop of the obtained results, which is generally not observed to such an extent in the case of the other ODFs. The information carried by individual ODFs overlaps in a non-trivial way, which may be observed e.g. on the basis of the PHK onset detection function (Eqn. (3)): the set WPD+PHK+CD performs definitely better than WPD+CD, while SF+WPD+PHK is actually worse than SF+WPD, indicating (perhaps) the need to increase the number of hidden neurons in the network structure. The problem complexity also partially stems from the heterogeneity of the sound sources in our database. Replacing one onset detection function with another may help some instrument types or a music genre, while degrading the results for others. An example shown in Fig. 8 reveals that changing PHK to WPD in 3-ODF configuration, although generally yields inferior results, for piano recordings leads to some enhancement (best seen for the piano+vocal recording in the *classic2* set). A repository containing more detailed results is available (Stasiak, 2015) and may be used for further comparison and analysis.

The obtained results are comparable to state-of-the-art solutions (cf. the F-measure results reported in MIREX (2013): eleven algorithms, median: 0.8025, max: 0.8727). The results reported in the literature for this particular dataset (Daudet *et al*., 2004) are also similar, including, e.g., the recall value of ca 80% for the EER point in the work of Alonso *et al*. (2005) or the values of the F-measure for several instruments (violin: 87%, trumpet: 89%, piano: 98%) obtained on the extended Leveau database by Lee and Kuo (2006).

5. Conclusions

In this work a solution of the onset detection problem in musical signals, based on a feed-forward multi-layer non-linear neural network, was presented. Unlike many other approaches based on direct analysis of the spectrum, the intermediate representation built upon classical onset detection functions was applied. In this way the input data are already presented in domain-relevant form, which allows simplifying the construction of the neural network and the training process.

The obtained results are comparable to state-of-the-art solutions. It should be noted that several improvements may be easily introduced into the proposed method, including, e.g., application of more perceptually-motivated onset detection functions (and a different number of ODFs, if necessary), controlling the generalization error and modification of the stop criteria for the training process and, eventually, preliminary automatic classification of the recordings with respect to the type of instruments. This last operation may allow us to use several networks trained for different onset types, similarly as in the presented second testing scheme.

Concluding, a decided advantage of the presented solution is its relative simplicity and extensibility. In

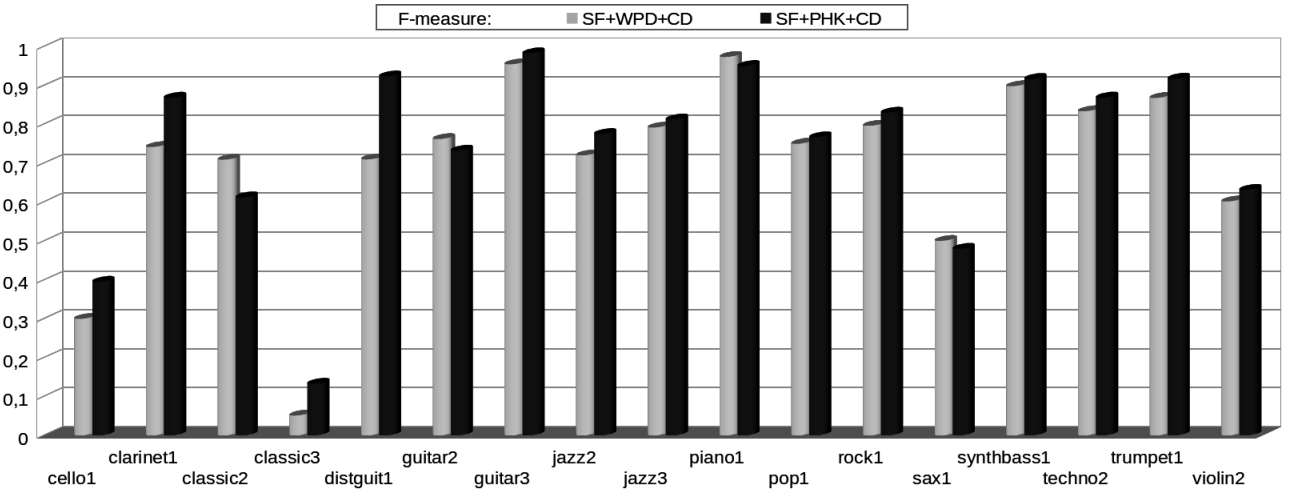


Fig. 8. Comparison of the F-measure value obtained for two different ODF subgroups.

future work, more training data will be used to obtain more reliable models, leading to further improvements of the results.

References

Alonso, M., Richard, G. and David, B. (2005). Extracting note onsets from musical recordings, *Proceedings of the IEEE International Conference on Multimedia and Expo 2005, Amsterdam, The Netherlands*, pp. 1–4.

Bartkowiak, M. and Januszkiewicz, Ł. (2012). Hybrid sinusoidal modeling of music with near transparent audio quality, *Proceedings of the Joint AES/IEEE Conference NTAV-SPA, Łódź, Poland*, pp. 91–96.

Bello, J., Daudet, L., Abdullah, S., Duxbury, C., Davies, M. and Sandler, M. (2005). A tutorial on onset detection in music signals, *IEEE Transactions on Speech and Audio Processing* **13**(5): 1035–1047.

Bello, P. and Sandler, M. (2003). Phase-based note onset detection for music signals, *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing ICASSP, Hong Kong*, Vol. 5, pp. 441–444.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*, Oxford University Press, New York, NY.

Böck, S., Arzt, A., Krebs, F. and Schedl, M. (2012). Online real-time onset detection with recurrent neural networks, *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx 2012), York, UK,* pp. 1–4.

Collins, N. (2005). A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions, *Proceedings of the AES 118th International Convention, Barcelona, Spain*, pp. 28–31.

Daudet, L., Richard, G. and Leveau, P. (2004). Methodology and tools for the evaluation of automatic onset detection algorithms in music, *5th International Conference on Music Information Retrieval, ISMIR 2004, Barcelona, Spain*, pp. 72–75.

Davy, M. and Godsill, S.J. (2002). Detection of abrupt spectral changes using support vector machines: An application to audio signal segmentation, *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2002, Orlando, FL, USA*, pp. 1313–1316.

Dixon, S. (2006). Onset detection revisited, *Proceedings of the International Conference on Digital Audio Effects (DAFx-06), Montreal, Quebec, Canada*, pp. 133–137.

Duxbury, C., Bello, J., Davies, M. and Sandler, M. (2003). Complex domain onset detection for musical signals, *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03), London, UK*, pp. 1–4.

Eyben, F., Böck, S., Schuller, B. and Graves, A. (2010). Universal onset detection with bidirectional long shortterm memory, *Neural Networks, 11 th International Society for Music Information Retrieval Conference (ISMIR 2010), Utrecht, The Netherlands*, pp. 589–594.

Huang, S., Wang, L., Hu, S., Jiang, H. and Xu, B. (2008). Query by humming via multiscale transportation distance in random query occurrence context, *IEEE International Conference on Multimedia and Expo, ICME 2008, Hannover, Germany*, pp. 1225–1228.

Lacoste, A. and Eck, D. (2007). A supervised classification algorithm for note onset detection, *EURASIP Journal of Advanced Signal Processing* **2007**: 153–153.

Laroche, J. (2003). Efficient tempo and beat tracking in audio recordings, *Journal of the Audio Engineering Society* **51**(4): 226–233.

Lee, W.-C. and Kuo, C.-C. (2006). Musical onset detection based on adaptive linear prediction, *IEEE International Conference on Multimedia and Expo, ICME 2006, Toronto, Ontario, Canada*, pp. 957–960.

Lerch, A. (2012). *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*, Wiley/IEEE Press, Hoboken, NJ.

MIREX (2013). Audio onset detection results in Music Information Retrieval Evaluation eXchange MIREX, 2013,

http://nema.lis.illinois.edu/nema_out/mirex2013/results/aod/summary.html.

Peeters, G. (2005). Time variable tempo detection and beat marking, *Proceedings of the International Computer Music Conference, ICMC 2005, Barcelona, Spain*, pp. 1–4.

Quintela, N.D., Giménez, A.P. and Guijarro, S.T. (2009). A comparison of score-level fusion rules for onset detection in music signals, *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR09, Kobe, Japan*, pp. 117–121.

Rabenstein, R. and Petrausch, S. (2008). Block-based physical modeling with applications in musical acoustics, *International Journal of Applied Mathematics and Computer Science* **18**(3): 295–305, DOI: 10.2478/v10006-008-0027-6.

Repp, B.H. (1996). Patterns of note onset asynchronies in expressive piano performance, *Journal of the Acoustical Society of America* **100**(6): 3917–3932.

Schlüter, J. and Böck, S. (2014). Improved musical onset detection with convolutional neural networks, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014), Florence, Italy,* pp. 6979–6983.

Stasiak, B. (2015). Results repository, http://ics.p.lodz.pl/~basta/NN_MULTI_ODF_FUSION/Stasiak_OnsetDB.zip.

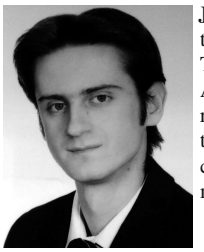
Tian, M., Fazekas, G., Black, D.A.A. and Sandler, M. (2014). Design and evaluation of onset detectors using different fusion policies, *15th International Society of Music Information Retrieval (ISMIR) Conference, ISMIR 2014, Taipei, Taiwan*, pp. 631–636.

Typke, R., Wiering, F. and Veltkamp, R.C. (2007). Transportation distances and human perception of melodic similarity, *Musicae Scientiae* **11**(1): 153–181.

Yin, J., Wang, Y. and Hsu, D. (2005). Digital violin tutor: An integrated system for beginning violin learners, *in* H. Zhang *et al.* (Eds.), *ACM Multimedia*, ACM, New York, NY, pp. 976–985.

Zhang, B. and Wang, Y. (2009). Automatic music transcription using audio-visual fusion for violin practice in home environment, *Technical Report TRA7/09*, National University of Singapore, Singapore.

Bartłomiej Stasiak received the M.Sc. degree in music from the Music Academy of Łódź in 2001, the M.Sc. degree in computer science from the Łódź University of Technology in 2004 and the Ph.D. degree in computer science from the Gdańsk University of Technology in 2010. He is an assistant professor at the Institute of Information Technology at the Łódź University of Technology. His research interests include artificial intelligence applications in image recognition, sound signal processing and music information retrieval.



Jędrzej Mońko received the M.Sc. degree from the Łódź University of Technology, Faculty of Technical Physics, Information Technology and Applied Mathematics (2012), where he is currently a Ph.D. student at the Institute of Information Technology. His research interests include computer graphics, multimedia and music information retrieval.

Adam Niewiadomski received the M.Sc. degree from the Łódź University of Technology, Poland, in 1998, and the Ph.D. and D.Sc. degrees from the Polish Academy of Sciences, Warsaw, in 2001 and 2009, respectively, all in computer science. He is currently an associate professor with the Institute of Information Technology, Łódź University of Technology. He is the author or a coauthor of more than 70 technical papers. His research interests include methods of computational intelligence, fuzzy representations of information, automated theorem proving, extensions of fuzzy sets, and e-learning.

Received: 25 June 2014
Revised: 13 May 2015