DESIGN AND EVALUATION OF ONSET DETECTORS USING DIFFERENT FUSION POLICIES

Mi Tian, György Fazekas, Dawn A. A. Black, Mark Sandler
Centre for Digital Music, Queen Mary University of London
{m.tian, g.fazekas, dawn.black, mark.sandler}@qmul.ac.uk

ABSTRACT

Note onset detection is one of the most investigated tasks in Music Information Retrieval (MIR) and various detection methods have been proposed in previous research. The primary aim of this paper is to investigate different fusion policies to combine existing onset detectors, thus achieving better results. Existing algorithms are fused using three strategies, first by combining different algorithms, second, by using the linear combination of detection functions, and third, by using a late decision fusion approach. Large scale evaluation was carried out on two published datasets and a new percussion database composed of Chinese traditional instrument samples. An exhaustive search through the parameter space was used enabling a systematic analysis of the impact of each parameter, as well as reporting the most generally applicable parameter settings for the onset detectors and the fusion. We demonstrate improved results attributed to both fusion and the optimised parameter settings.

1. INTRODUCTION

The automatic detection of onset events is an essential part in many music signal analysis schemes and has various applications in content-based music processing. Different approaches have been investigated for onset detection in recent years [1,2]. As the main contribution of this paper, we present new onset detectors using different fusion policies, with improved detection rates relying on recent research in the MIR community. We also investigate different configurations of onset detection and fusion parameters, aiming to provide a reference for configuring onset detection systems.

The focus of ongoing onset detection work is typically targeting Western musical instruments. Apart from using two published datasets, a new database is incorporated into our evaluation, collecting percussion ensembles of Jingju, also denoted as Peking Opera or Beijing Opera, a major genre of Chinese traditional music [1]. By including this dataset, we aim at increasing the diversity of instrument categories in the evaluation of onset detectors, as well as extending the research to include non-Western music types.

The goal of this paper can be summarised as follows: *i)* to evaluate fusion methods in comparison with the baseline algorithms, as well as a state-of-the-art method [2] ; *ii)* to investigate which fusion policies and which pair-wise combinations of onset detectors yield the most improvement over standard techniques; *iii)* to find the best performing configurations by searching through the multi-dimensional parameter space, hence identifying emerging patterns in the performances of different parameter settings, showing good results across different datasets; *iv)* to investigate the performance difference in Western and non-Western percussive instrument datasets.

In the next section, we present a review of related work. Descriptions of the datasets used in this experiment are given in Section 3. In Section 4, we introduce different fusion strategies. Relevant post-processing and peak-picking procedures, as well as the parameter search process will be discussed in Section 5. Section 6 presents the results, with a detailed analysis and discussion of the performance of the fusion methods. Finally, the last section summarises our findings and provides directions for future work.

2. RELATED WORK

Many onset detection algorithms and systems have been proposed in recent years. Common approaches using energy or phase information derived from the input signal include the high frequency content (HFC) and complex domain (CD) methods. See [1,6] for detailed reviews and [9] for further improvements. Pitch contours and harmonicity information can also be indicators for onset events [7]. These methods shows some superiority over energy based ones in case of soft onsets.

Onset detection systems using machine learning techniques have also been gaining popularity in recent years [3]. The winner of MIREX 2013 audio onset detection task utilises convolutional neural networks to classify and distinguish onsets from non-onset events in the spectrogram [13]. The data-driven nature of these methods makes the

[1] http://en.wikipedia.org/wiki/Peking_opera

[2] Machine learning-based methods are excluded from this study to limit the scope of our work.

[3] http://www.music-ir.org/mirex/wiki/2013:Audio_Onset_Detection

detection less dependent on onset types, though a computationally expensive training process is required. A promising approach for onset detection lies in the fusion of multiple detection methods. Zhou et al. proposed a system integrating two detection methods selected according to properties of the target onsets [17]. In [10], pitch, energy and phase information are considered in parallel for the detection of pitched onsets. Another fusion strategy is to combine peak score information to form new estimations of the onset events [8]. Albeit fusion has been used in previous work, there is a lack of systematic evaluation of fusion strategies and applications in the current literature. This paper focusses on the assessment of different fusion policies, from feature-level and detection function-level fusion to higher level decision fusion.

The success of an onset detection algorithm largely depends on the signal processing methods used to extract salient features from the audio that emphasise the features characterising onset events as well as smoothing the noise in the detection function. Various signal processing techniques have been introduced in recent studies, such as vibrato suppression [3] and adaptive thresholding [1]. In [14], adaptive whitening is presented where each STFT bins magnitude is divided by the an average peak for that bin accumulated over time. This paper also investigates the performances of some commonly used signal processing modules within onset detection systems.

3. DATASETS

In this study, we use two previously released evaluation datasets and a newly created one. The first published dataset comes from [1], containing 23 audio tracks with a total duration of 190 seconds and having 1058 onsets. These are classified into four groups: pitched non-percussive (PNP), e.g. bowed strings, 93 onsets, pitched percussive (PP), e.g. piano, 482 onsets [4], non-pitched percussive (NPP), e.g. drums, 212 onsets, and complex mixtures (CM), e.g. pop singing music, 271 onsets. The second set comes from [2] which is composed of 30 samples [5] of 10 second audio tracks, containing 1559 onsets in total, covering also four categories: PNP (233 onsets in total), PP (152 onsets), NPP (115 onsets), CM (1059 onsets). The use of these datasets enables us to test the algorithms on a range of different instruments and onset types, and provides for direct comparison with published work. The combined dataset used in the evaluation of our work is composed of these two sets.

The third dataset consists of recordings of the four major percussion instruments in Jingju: bangu (clapper- drum), daluo (gong-1), naobo (cymbals), and xiaoluo (gong-2). The samples are manually mixed using individual recordings of these instruments with possibly simultaneous onsets to closely reproduce real world conditions. See [15] for more details on the instrument types and the dataset. This dataset includes 10 samples of 30-second excerpts with 732 onsets. We also use NPP onsets from the first two datasets to form the fourth one, providing a direct comparison with the Chinese NPP instruments. All stimuli are mono signals sampled at 44.1kHz [6] and 16 bits per sample, having 3349 onsets in total.

4. FUSION EXPERIMENT

The aim of information fusion is to merge information from heterogeneous sources to reduce uncertainty of inferences [11]. In our study, six spectral-based onset detection algorithms are considered as baselines for fusion: high frequency content (HFC), spectral difference (SD) complex domain (CD), broadband energy rise (BER), phase deviation (PD), outlined in [1], and SuperFlux (SF) from recent work [4]. We also developed and included in the fusion a method based on Linear Predictive Coding [12], where the LPC coefficients are computed using the Levinson-Durbin recursion, and the onset detection function is derived from the LPC error signal.

Three fusion policies are used in our experiments: *i)* feature-level fusion, *ii)* fusion using the linear combination of detection functions and *iii)* decision fusion by selecting and merging onset candidates. All pairwise combination of the baseline algorithms are amenable for the latter two fusion policies. However, not all algorithms can be meaningfully combined using feature-level fusion. For example CD can be considered as an existing combination of SD and PD, therefore combining CD with either of these two at a feature level is not sensible. In this study, 10 feature-level fusion, 13 linear combination based fusion and 15 decision fusion based methods are tested. These are compared to the 7 original methods, giving us 45 detectors in total. In the following, we describe specific fusion policies. We assume familiarity with onset detection principles and restrain from describing these details, please see [1] for a tutorial.

4.1 Feature-level Fusion

In feature-level fusion, multiple algorithms are combined to compute fused features. For conciseness, we provide only one example combining BER and SF, denoted BERSF, utilising the vibrato suppression capability of SF [4] for detecting soft onsets, as well as the good performance of BER for detecting percussive onsets with sharp energy bursts [1]. Here, we use the BER to mask the SF detection function as described by Equation (1). In essence, SF is used directly when there is evidence for a sharp energy rise, otherwise it is further smoothed using a median filter.

$$ODF(n) = \begin{cases} SF(n) & \text{if } BER(n) > \gamma \\ \lambda(\overline{SF(n)}) & \text{otherwise,} \end{cases} \qquad (1)$$

where γ is an experimentally defined threshold, λ is a weighting constant set to 0.9 and $\overline{SF(n)}$ is the median filtered detection function with a window size of 3 frames.

[4] A 7-onset discrepancy(482 instead of 489) from the reference paper is reported by the original author due to revisions of annotations.

[5] Only a subset of this dataset presented in the original paper is received from the author for the evaluation in this paper.

[6] Some audio files were upsampled to obtain a uniform dataset.

4.2 Linear Combination of Detection Functions

In this method, two time aligned detection functions are used and their weighted linear combination is computed to form a new detection function as shown in Equation 2:

$$ODF(n) = wODF_1(n) + (1 - w)ODF_2(n), \qquad (2)$$

where ODF_1 and ODF_2 are two normalised detection functions and w is a weighting coefficient $(0 \leq w \leq 1)$.

4.3 Decision Fusion

This fusion method operates at a later stage and combines prior decisions of two detectors. Post-processing and peak picking are applied separately yielding two lists of onset candidates. Onsets from the two lists occurring within a fixed temporal tolerance window will be merged and accepted. Let TS_1 and TS_2 be the lists of onset locations given by two different detectors, *i* and *j* be indexes of onsets in the candidate lists and δ the tolerance time window. The final onset locations are generated using the fusion strategy described by Algorithm 1.

Algorithm 1 Onset decision fusion

```
procedure DecisionFusion($TS_1, TS_2$)
    $I, J \leftarrow 0 : len(TS_1) - 1, 0 : len(TS_2) - 1$
    $TS \leftarrow$ empty list
    for all $i, j$ in $product(I, J)$ do
        if $abs(TS_1[i] - TS_2[j]) < \delta$ then
            insert sorted: $TS \leftarrow mean(TS_1[i], TS_2[j])$
    return $TS$
```

5. PEAK PICKING AND PARAMETER SEARCH

5.1 Smoothing and Thresholding

Post-processing is an optional stage to reduce noise that interferes with the selection of maxima in the detection function. In this study, three post-processing blocks are used: *i)* DC removal and normalisation, *ii)* zero-phase low-pass filtering and *iii)* adaptive thresholding. In conventional normalisation, data is scaled using a fixed constant. Here we use a normalisation coefficient computed by weighting the input exponentially. After removing constant offsets, the detection function is normalised using the coefficient *AlphaNorm* calculated by Equation (3):

$$AlphaNorm = \left(\frac{\sum_n |ODF(n)|^\alpha}{len(ODF)} \right)^{\frac{1}{\alpha}} \qquad (3)$$

A low-pass filter is applied to the detection function to reduce noise. To avoid introducing delays, a zero phase filter is employed at this stage. Finally, adaptive thresholding using a moving median filter is applied following Bello [1], to avoid the common pitfalls of using a fixed threshold for peak picking.

5.2 Peak Picking

5.2.1 Polynomial Fitting

The use of polynomial fitting allows for assessing the shape and magnitude of peaks separately. Here we fit a second-degree polynomial on the detection function around local maxima using a least squares method, following the QM Vamp Plugins [7]. The coefficients a and c of the quadratic equation $y = ax^2 + bx + c$ are used to detect both sharper peaks, under the condition $a > th_a$, and peaks with a higher magnitude, when $c > th_c$. The corresponding thresholds are computed from a single sensitivity parameter called $threshold$ using $th_a = (100 - threshold)/1000$ for the quadratic term and $th_c = (100 - threshold)/1500$ for the constant term. The linear term b can be ignored.

5.2.2 Backtracking

In case of many musical instruments, onsets have longer transients without a sharp burst of energy rise. This may cause energy based detection functions to exhibit peaks after the perceived onset locations. Vos and Rasch conclude that onsets are perceived when the envelope reaches a level of roughly 6-15 dB below the maximum level of the tones [16]. Using this rationale, we trace the onset locations from the detected peak position back to a hypothesised earlier "perceived" location. The backtracking procedure is based on measuring relative differences in the detection function, as illustrated by Algorithm 2, where θ is the threshold used as a stopping condition. We use the implementation available in the QM Vamp Plugins.

Algorithm 2 Backtracking

```
Require: idx: index of a peak location in the ODF
procedure Backtracking($idx, ODF, \theta$)
    $\delta, \gamma \leftarrow 0$
    while $idx > 1$ do
        $\delta \leftarrow ODF[idx] - ODF[idx - 1]$
        if $\delta < \gamma * \theta$ then
            break
        $idx \leftarrow idx - 1$
        $\gamma \leftarrow \delta$
    return $idx$
```

5.3 Parameter Search

An exhaustive search is carried out to find the configurations in the parameter space yielding the best detection rates. The following parameters and settings, related to the onset detection and fusion stages, are evaluated: *i)* adaptive whitening (**wht**) on/off; *ii)* detection sensitivity (threshold), ranging from 0.1 to 1.0 with an increment of 0.1; *iii)* backtracking threshold ($\boldsymbol{\theta}$), ranging from 0.4 to 2.4 with 8 equal subdivisions (the upper bound is set to an empirical value 2.4 in the experiment since the tracking will not go beyond the previous valley); *iv)* linear combination coefficient ($\boldsymbol{w}$), ranging from 0.0 to 1.0 with an increment of 0.1; *v)* tolerance window length ($\boldsymbol{\delta}$) for decision fusion, ranging from 0.01 to 0.05 (in second) having 8 subdivisions. This gives a 5-dimensional space and all combinations of all possible values described above are evaluated. This results in 180 configurations in case of standard detectors and feature-level fusion, 1980 in case of linear fusion and 1620 for decision fusion. The configurations are described

[7] http://www.vamp-plugins.org

using the Vamp Plugin Ontology [8] and the resulting RDF files are used by Sonic Annotator [5] to configure the detectors. The test result will thus give us not only the overall performance of each onset detector, but also uncover their strengths and limitations across different datasets and parameter settings.

6. EVALUATION AND RESULTS

6.1 Analysis of Overall Performance

Figure 1 provides an overview of the results, showing the F-measure for the top 12 detectors in our study [9]. Detectors are ranked by the median showing the overall performance increase due to fusion across the entire range of parameter settings. Due to space limitations, only a subset of the results are reported in this paper. The complete result set for all tested detectors under all configurations on different datasets is available online [10], together with Vamp plugins of all tested onset detectors. The names of the fusion algorithms come from the abbreviations of the constituent methods, while the numbers represent the fusion policy: *0: feature-level fusion*, *1: linear combination of detection functions* and *2: decision fusion*.

CDSF-1 yields improved F-measure for the combined dataset by 3.06% and 6.14% compared to the two original methods SF and CD respectively. Smaller interquartile ranges (IQRs) observed in case of CD, SD and HFC based methods show they have less dependency on the configuration. BERSF-2 and BERSF-1 vary the most in performance, also reflected from their IQRs. In case of BERSF-2, the best performance is obtained using the widest considered tolerance window (0.05s), with modest sensitivity (40%). However, decreasing the tolerance window size has an adverse effect on the performance, yielding one of the lowest detection rates caused by the significant drop of recall. In case of BERSF-1, a big discrepancy between the best and worst performing configurations can be observed. This is partly because the highest sensitivity setting has a negative effect on SF causing very low precision.

Table 1 shows the results ranked by F-measure, precision and recall with corresponding standard deviations for the ten best detectors as well as all baseline methods. Standard deviations are computed over the results for all configurations in each dataset. SF is ranked in the best performing ten, thus it is excluded from the baseline. Nine out of the top ten detectors are fusion methods. CDSF-1 performs the best for all datasets (including CHN-NPP and WES-NPP that are not listed in the table) while BERSF yields the second best performance in the combined, WES-NPP and JPB datasets. Corresponding parameter settings for the combined dataset are given in Table 2.

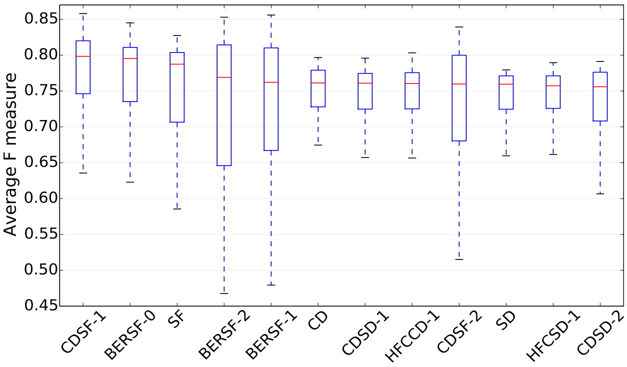

Figure 1. F-meaure of all configurations for the top 12 detectors. (Min, first and third quartile and max value of the data are represented by the bottom bar of the whiskers, bottom and upper borders of the boxes and upper bar of the whiskers respectively. Median is shown by the red line)

method	threshold	θ	wht	w	δ (s)
CDSF-1	10.0	2.15	off	0.20	n/a
BERSF-1	10.0	2.40	off	0.30	n/a
BERSF-2	40.0	2.15	off	n/a	0.05
BERSF-0	30.0	2.40	off	n/a	n/a
CDSF-2	50.0	2.40	off	n/a	0.05
SF	20.0	2.40	off	n/a	n/a
CDBER-1	10.0	2.40	off	0.50	n/a
BERSD-1	10.0	2.40	off	0.60	n/a
HFCCD-1	20.0	1.15	off	0.50	n/a
CDBER-2	50.0	1.15	off	n/a	0.05
mean	25.90	2.100	-	0.4200	0.05
std	15.01	0.4848	-	0.1470	0.00
median	20.0	2.15	-	0.50	0.05
mode	10.0	2.40	off	0.50	0.05

Table 2. Parameter settings for the ten best performing detectors, **threshold**: overall detection sensitivity; $\boldsymbol{\theta}$: backtracking threshold; **wht**: adaptive whitening; $\boldsymbol{w}$: linear combination coefficient; $\boldsymbol{\delta}$: tolerance window size.

Fusion policies may perform differently in the evaluation. In case of feature-level fusion, we compared how combined methods score relative to their constituents. The performances vary between datasets, with only HFCBER-0 outperforming both HFC and BER on the combined and SB datasets in terms of mean F-measure. However, five perform better than their two constitutes on JPB, two on CHN-NPP and five on WES-NPP dataset (these results are published online). A more detailed analysis of these performance differences constitutes future work.

When comparing linear fusion of detection functions with decision fusion, the former performs better across all datasets in all but one cases, the fusion of HFC and BER. Even in this case, linear fusion yields close performance in terms of mean F-measure. Interesting observations also emerge for particular methods on certain datasets. The linear fusion based detectors involving LPC and PD (SDPD-1 and LPCPD-1) show better performances in the case of the CHN-NPP dataset compared to their performances on other datasets as well those given by their constituent methods (please see table online). Further analysis, for instance, by looking at statistical significance of these observations is required to identify relevant instrument properties.

When comparing BERSF-2, CDSF-2 and CDBER-2 to the other detectors in Table 1, notably higher standard deviations in *recall* and *F-measure* are shown, indicating this

[8] `http://www.omras2.org/VampOntology`

[9] Due to different post-processing stages, the results reported here may diverge from previously published results.

[10] `http://isophonics.net/onset-fusion`

method	F (combined)	P (combined)	R (combined)	F (sb)	P (sb)	R (sb)	F (jpb)	P (jpb)	R (jpb)
CDSF-1	**0.8580** 0.0613	**0.9054** 0.1195	0.8153 **0.0609**	**0.8194** 0.0598	0.8455 0.1165	0.7949 **0.0681**	**0.9286** 0.0649	**0.9748** 0.1241	0.8865 **0.0525**
BERSF-1	0.8559 0.0941	0.8857 0.1363	**0.8280** 0.0866	0.8126 0.0961	0.8191 0.1306	0.8062 0.0988	0.9283 0.0925	0.9718 0.1463	**0.8885** 0.0710
BERSF-2	0.8528 0.1684	0.8901 0.1411	0.8186 0.2028	0.8088 0.1677	**0.8729** 0.1470	0.7536 0.2055	0.9230 0.1724	0.9637 0.1310	0.8856 0.2011
BERSF-0	0.8451 0.0722	0.8638 0.1200	0.8272 0.0701	0.8025 0.0723	0.8185 **0.1134**	0.7870 0.0744	0.9175 0.0747	0.9712 0.1322	0.8694 0.0658
CDSF-2	0.8392 0.1537	0.8970 **0.1129**	0.7884 0.1855	0.7892 0.1758	0.8336 0.1251	0.7493 0.2014	0.9165 0.1344	0.9642 **0.1001**	0.8732 0.1690
SF	0.8274 0.0719	0.8313 0.1209	0.8234 0.0657	0.8126 0.0744	0.8191 0.1241	**0.8063** 0.0737	0.8488 0.0704	0.8290 0.1177	0.8694 0.0558
CDBER-1	0.8145 0.0809	0.8210 0.1276	0.8080 0.0792	0.7877 0.0829	0.7972 0.1295	0.7785 0.0893	0.8560 0.0793	0.8678 0.1253	0.8446 0.0667
BERSD-1	0.8073 0.0792	0.8163 0.1311	0.7986 0.0812	0.7843 0.0828	0.7985 0.1358	0.7707 0.0915	0.8420 0.0756	0.8310 0.1252	0.8532 0.0685
HFCCD-1	0.8032 **0.0472**	0.8512 0.1179	0.7603 0.0734	0.7802 **0.0448**	0.8387 0.1239	0.7293 0.0765	0.8416 **0.0511**	0.8376 0.1101	0.8456 0.0705
CDBER-2	0.7967 0.2231	0.8423 0.1404	0.7558 0.2398	0.7605 0.2279	0.8140 0.1607	0.7138 0.2384	0.8498 0.2291	0.8853 0.1273	0.8170 0.2494
CD	0.7966 0.0492	0.8509 0.1164	0.7489 0.0672	0.7692 0.0467	0.8361 0.1191	0.7123 0.0709	0.8320 0.0535	0.8692 0.1128	0.7979 0.0636
BER	0.7883 0.0942	0.7776 0.1184	0.7994 0.1001	0.7626 0.0974	0.7521 0.1166	0.7138 0.1119	0.8254 0.0920	0.7968 0.1226	0.8561 0.0851
SD	0.7795 0.0466	0.8354 0.1269	0.7305 0.0733	0.7604 0.0450	0.8311 0.1326	0.7009 0.0785	0.8210 0.0491	0.8202 0.1190	0.8217 0.0676
HFC	0.7712 0.0412	0.8011 0.1225	0.7436 0.0898	0.7411 0.0375	0.7818 0.1291	0.7044 0.0844	0.8159 0.0496	0.8082 0.1138	0.8236 0.1002
LPC	0.7496 0.0658	0.7671 0.1103	0.7330 0.1061	0.7243 0.0657	0.7494 0.1069	0.7009 0.1019	0.7913 0.0662	0.8041 0.1164	0.7788 0.1118
PD	0.6537 0.1084	0.5775 0.1008	0.7530 0.2235	0.6143 0.1093	0.5230 0.0688	0.7308 0.2302	0.7114 0.1115	0.6513 0.1536	0.7836 0.2158

Table 1. F-measure (**F**), Precision (**P**) and Recall (**R**) for dataset **combined**, **SB**, **JPB** for detectors under best performing configurations from the parameter search, with corresponding standard deviations over different configurations.

statistic	Combined	SB	JPB	CHN-NPP	WES-NPP
mean	0.7731	0.7438	0.8183	0.8527	0.8358
std	0.0587	0.0579	0.0628	**0.1206**	0.0641
median	0.7818	0.7595	0.8226	0.8956	0.8580

Table 3. Statistics for F-measure of the ten detectors with their best performances from Table 1 for different datasets

fusion policy is more sensitive to the choice of parameters. A possible improvement in this fusion policy would be to make the size of the tolerance window dependent on the magnitude of relevant peaks of the detection functions.

The results also vary across different datasets. Table 3 summarises F-measure statistics computed over the detectors listed in Table 1 at their best setting for each datasets used in this paper. In comparison with SB, the JPB dataset exhibits higher F-measure. This dataset has larger diversity in terms of the length of tracks and the level of complexity, while the SB dataset mainly consists of complex mixture (CM) onsets type. Both the Chinese and Western NPP onset class provides noticeably higher detection rate compared to the mix-typed datasets. Though the CHN-NPP set shows the largest standard deviation, suggesting a greater variation in performance between the different detectors for these instruments. Apart from aiming at optimal overall detection results, it is also useful to consider when and how a certain onset detector exhibits the best performance, which constitutes future work.

6.2 Parameter Specifications

For general datasets a low detection sensitivity value is favourable, which is supported by the fact that 30 out of the 45 tested methods yield the best performances with a sensitivity lower than 50% (see online). In 23 out of all cases, the value of the backtracking threshold was the highest considered in our study (2.4) when the detectors yield the best performances for the combined dataset, and it was unanimously at a high value for all other datasets including the percussive ones. This suggests that in many cases, the perceived onset will be better characterised by the valley of the detection function prior to the detected peak. Note that even at a higher threshold, the onset location would not be traced back further than the valley preceding the peak detected in our algorithm. An interesting direction for future work would thus be, given this observation, to take into account the properties of human perception.

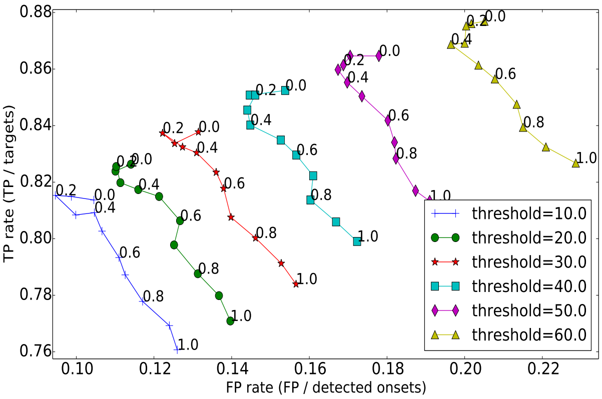

Figure 2. Performances of **CDSF-1** onset detector under different $\boldsymbol{w}$ (labelled in each curve) and **threshold** (annotated in the side box) settings

Adaptive whitening had to be turned off for the majority of detectors to provide good performance for all datasets. This indicates that the method does not improve onset detection performance in general, although it is available in most onset detectors in the Vamp plugin library. The value of the tolerance window was always 0.05s for best performance in our study, suggesting that the temporal precision of the different detectors varies significantly, which requires a fairly wide decision horizon for successful combination.

Figure 2 shows how two parameters influence the performance of the onset detector CDSF-1. The figure illustrates the true positive rate (i.e., correct detections relative to the number of target onsets) and false positive rate (i.e., false detections relative to the number of detected onsets) and better performance is indicated by the curve shifting upwards and leftwards. All parameters except the *linear combination coefficient* $(\boldsymbol{w})$ and *detection sensitivity* *(threshold)* are fixed at their optimal values. We can observe that the value of the linear combination coefficient is around 0.2 for best performance. This suggests that the detector works the best when taking the majority of the contribution from SF. With the *threshold* increasing from 10.0% to 60.0%, the true positive rate is increasing at the cost of picking more false onsets, thus a lower sensitivity is preferred in this case. Poorest performance in case of the linear fusion policy occurs in general when the linear combination coefficient overly favours one constituent detector, or the sensitivity (threshold) is too high and the backtracking threshold (θ) is at its lowest value.

7. CONCLUSION AND FUTURE WORK

In this work, we applied several fusion techniques to aid the music onset detection task. Different fusion policies were tested and compared to their constituent methods, including the state-of-the-art SuperFlux method. A large scale evaluation was performed on two published datasets showing improvements as a result of fusion, without extra computational cost, or the need for a large amount of training data as in the case of machine learning based methods. A parameter search was used to find the optimal settings for each detector to yield the best performance.

We found that some of the best performing configurations do not match the default settings of some previously published algorithms. This suggests that in some cases, better performance can be achieved just by finding better settings which work best overall for a given type of audio even without changing the algorithms.

In future work, a possible improvement in case of late decision fusion is to take the magnitude of the peaks into account when combining detected onsets, essentially treating the value as an estimation confidence. We will investigate the dependency of the selection of onset detectors on the type and the quality of the input music signal. We also intend to carry out more rigorous statistical analyses with significance tests for the reported results. More parameters could be included in the search to study their strengths as well as how they influence each other under different configurations. Another interesting direction is to incorporate more Non-Western music types as detection target and design algorithms using instrument specific priors.

8. REFERENCES

[1] J.P. Bello, L. Daudet, S. Abdallan, C. Duxbury, and M. Davies. A tutorial on onset detection in music signals. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 13, 2005.

[2] S. Böck, F. Krebs, and M. Schedl. Evaluating the online capabilities of onset detection methods. In *Proc. of the 13th Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2012.

[3] S. Böck and G. Widmer. Local group delay based vibrato and tremolo suppression for onset detection. In *Proc. of the 14th Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2013.

[4] S. Böck and G. Widmer. Maximum filter vibrato suppression for onset detection. In *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx)*, 2013.

[5] C. Cannam, M.O. Jewell, C. Rhodes, M. Sandler, and M. d'Inverno. Linked data and you: Bringing music research software into the semantic web. *Journal of New Music Research*, 2010.

[6] N. Collins. A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. In *Proc. of the 118th Convention of the Audio Engineering Society*, 2005.

[7] N. Collins. Using a pitch detector for onset detection. In *Proc. of the 6th Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2005.

[8] N. Degara-Quintela, A. Pena, and S. Torres-Guijarro. A comparison of score-level fusion rules for onset detection in music signals. In *Proc. of the 10th Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2009.

[9] S. Dixon. Onset detection revisited. In *Proc. of the 9 th Int. Conference on Digital Audio Effects (DAFx'06)*, 2006.

[10] A. Holzapfel, Y. Stylianou, A.C. Gedik, and B. Bozkurt. Three dimensions of pitched instrument onset detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[11] L.A. Klein. *Sensor and data fusion: a tool for information assessment and decision making*. SPIE, 2004.

[12] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4), 1975.

[13] J. Schlüter and S. Böck. Musical onset detection with convolutional neural networks. In *6th Int. Workshop on Machine Learning and Music (MML)*, 2013.

[14] D. Stowell and M. Plumbley. Adaptive whitening for improved real-time audio onset detection. In *Proceedings of the International Computer Music Conference (ICMC)*, 2007.

[15] M. Tian, A. Srinivasamurthy, M. Sandler, and X. Serra. A study of instrument-wise onset detection in beijing opera percussion ensembles. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[16] J. Vos and R. Rasch. The perceptual onset of musical tones. *Perception & Psychophysics*, 29(4), 1981.

[17] R. Zhou, M. Mattavellii, and G. Zoia. Music onset detection based on resonator time frequency image. *IEEE Transactions on Audio, Speech, and Language Processing*, 2008.