NOTE ONSET DETECTION USING RHYTHMIC STRUCTURE

Norberto Degara, Antonio Pena[*]

University of Vigo
Signal Theory and Communications Dept.
E.T.S.E. Telecomunicacion, Vigo, Spain

Matthew E. P. Davies, Mark D. Plumbley[†]

Queen Mary University of London
Centre for Digital Music
Mile End Road, London E1 4NS, UK

ABSTRACT

In this paper we explore the relationship between the temporal and rhythmic structure of musical audio signals. Using automatically extracted rhythmic structure we present a rhythmically-aware method to combine note onset detection techniques. Our method uses top-down knowledge of repetitions of musical events to improve detection performance by modelling the temporal distribution of onset locations. Results on a publicly available database demonstrate that using musical knowledge in this way can lead to significant improvements by reducing the number of missed and spurious detections.

Index Terms— Audio, music, onset detection, rhythm

1. INTRODUCTION

The task of recovering the start times of musical events from audio signals is known as *note onset detection* [1]. The successful extraction of note onset times enables the temporal segmentation of an audio signal at a meaningful time-scale. Within music information retrieval research, onset detection forms the basis of many higher level processing tasks, including beat tracking [2] and interactive musical accompaniment [3]. The standard approach for finding onset positions is a two stage process. First, a mid-level representation, often referred to as an *onset detection function* [1], is extracted from the audio signal. The aim of the onset detection function is to exhibit peaks at likely onset locations by measuring changes in the short term properties of the audio signal; for example: energy, high frequency content, or phase information. For a review of feature types see [1]. Once the onset detection function has been generated, the temporal locations of the note onsets can be recovered by applying a *peak-picking* algorithm.

A key challenge in onset detection is in finding features which can accurately capture different types of onsets. For example, *pitched non-percussive* onsets from a bowed violin and *non-pitched percussive* onsets from drum hits correspond to very different properties of the audio signal. While there has been moderate success in finding features that are applicable the widest range of signals possible, e.g. the complex spectral difference onset detection function [1], more recent approaches have looked to choose one of different types of features, e.g. the energy and pitch based approaches of Zhou et al [4]. Although the use of multiple features might appear an intuitive step, it adds complexity in terms of how best to fuse these information sources.

Beyond the selection of appropriate input features, a further limitation exists within existing work, related to the temporal structure of music. The temporal ordering of musical events and their repetition is central to our perception of rhythm. Therefore when seeking to find onset locations, making the assumption that musical events can occur at *any* time instant is musically naïve. In this sense, Grosche and Müller [5] have recently proposed an algorithm that exploits the local periodic structure of musical events.

In this paper we address the use of multiple features and the inclusion of musical knowledge towards the advancement of note onset detection. Our aim is not to present a new type of onset detection function per se, but to propose a novel strategy for combining these types of signals.

To contend with the different types of onset that may be present in the audio signal, we adopt a mixture of experts approach [6] for fusing the peak locations extracted from a set of onset detection functions. Through observing the distribution of inter-onset-intervals we can determine the likelihood of given onset locations based on their relationship with surrounding events. We incorporate this within our system as a rhythmic constraint in our fusion algorithm. In our evaluation, we demonstrate that our approach, i.e. the use of multiple experts fused using rhythmic structure, can lead to an increase in onset detection accuracy

The remainder of this paper is structured as follows. In Section 2 we present our system for fusing onsets using knowledge of rhythmic structure. In Section 3 we describe the evaluation metric and dataset used with results in Section 4. We present conclusions and future work in Section 5.

2. APPROACH

Our algorithm for fusing onsets using musical knowledge of rhythmic structure is split into several steps. We modify an existing state of the art onset detection system to give sub-band onset detection functions. We then fuse the peaks of each sub-band onset detection functions. Given this initial fusion, we extract an estimate of rhythmic structure and then implement a second peak fusion stage incorporating rhythmic knowledge. A block diagram is shown in Figure 1.

2.1. Sub-band Onset Detection

To best demonstrate the potential improvement our fusion method can provide, we apply our rhythmic fusion strategy to an existing state of the art onset detection technique. We choose the winning algorithm from the MIREX 2007 onset detection evaluation task[1], that of Zhou et al [4]. Their approach generates an energy-based onset

[*] {ndegara, apena}@gts.tsc.uvigo.es. ND and AP are supported by the Spanish MEC, ref. TEC2006-13883-C04-02, under the project AnClaS3 “Sound source separation for acoustic measurements”.

[†] {matthew.davies, mark.plumbley}@elec.qmul.ac.uk. MEPD and MDP are supported by EPSRC Grants EP/G007144/1, EP/E045235/1 and EP/E017614/1.

[1] www.music-ir.org/mirex/2007/index.php/Audio_Onset_Detection

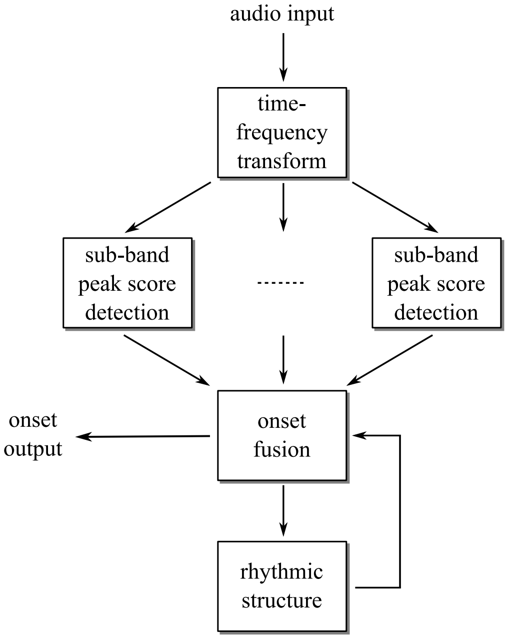

Fig. 1. Overview of Onset Fusion System.

detection function from a novel time frequency transform known as the *Resonator Time Frequency Image* (RTFI). The RTFI is calculated using 960 filters over B=8 musical octaves.

To generate multiple input features from Zhou's model that can be fused together, we calculate an individual energy based detection function for each of the 8 octaves (using 120 filters per octave), which we label $S_b(t)$, where $b = 1 \ldots B$ and t are the samples of the signal. Each sub-band has a temporal resolution 10ms per sample. For a complete description of the algorithm, see [4].

Our fusion method requires a set of sub-band peak scores $s_{b,j}$ for each associated time instant $t_{b,j}$. To obtain the time instants we employ a peak-picking algorithm [1] to each sub-band detection function $S_b(t)$. At this stage we do not wish to discard onsets and therefore we set the detection threshold parameter δ=0.

We extract a set of initial peak scores as the amplitude of each detection function at the peak time $t_{b,j}$, where the score of the m^{th} peak in the b^{th} sub-band is found as $s_{b,j} = S_b(t_{b,j})$.

To prevent any individual sub-band dominating in the temporal fusion of peaks, we normalise the influence of each sub-band, by mapping the peak scores, $s_{b,j}$ into the range [0,1] according to the empirical cumulative distribution function of sub-band peak scores, $\hat{F}_b$:

$$\tilde{s}_{b,j} = \hat{F}_b(s_{b,j}). \tag{1}$$

Then we order the whole set of peak scores, $\tilde{s}_{b,j}$, and peaks locations, $t_{b,j}$, over all sub-bands in time to give s_i and t_i respectively.

2.2. Onset Fusion with Temporal Constraints

In order to integrate the peak score information extracted from the multiple sub-band onset detection functions, we construct an objective function which is maximised according to two constraints: first, onsets should correspond to time instants where the sub-band detection functions show strong peaks and second, these peaks should be close together in time. The cost function that combines these two goals is:

$$C(\{t_i\}) = \sum_{i=1}^{N} s_i + \sum_{i=2}^{N} P(t_i - t_{i-1}, T_{\text{ref}}) \tag{2}$$

where $\{t_i\}$ is any set of N peak scores $\{t_1, ..., t_N\}$, $P(\Delta t, T_{\text{ref}})$ is a grouping penalty function and T_{ref} is a time reference that sets how fast the penalty term increases.

As in [7] the objective function $C(\{t_i\})$ can be assembled recursively using dynamic programming. We iteratively find the group of peaks $\{t_i\}$ that maximise the objective function at each time instant. Finally, to decide if a set of grouped peaks correspond to an onset or not, we extract the local maxima in $C(\{t_i\})$ within a window of 50 ms (assuming two consecutive onsets do not happen in this time). For the grouping penalty function we employ a squared-error function,

$$P(\Delta t, T_{\text{ref}}) = -\left(\frac{\Delta t}{T_{\text{ref}}}\right)^2 \tag{3}$$

which takes a value of -1 when $\Delta t = T_{\text{ref}}$ and becomes increasingly negative for larger time deviations between time peaks. Informal examination of the distribution of the peak times on the different bands showed that peaks that correspond to the same onset should not be more than 35 ms away from each other. Therefore we set T_{ref} = 35 ms in our experiments.

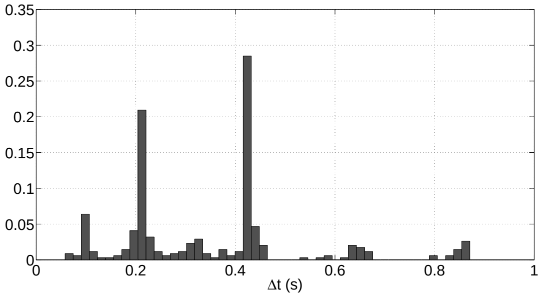

Fig. 2. Probability density estimate of the inter-onset-intervals extracted from the first onset fusion stage.

2.3. Finding Rhythmic Structure

To exploit the idea that note events are not uniformly distributed in time, we extract a set of onsets obtained from the temporal constrained fusion described in Section 2.2, and use this information to estimate the underlying rhythmic structure present in the input signal. To extract onsets from $C(\{t_i\})$ in (2), a large threshold ($\delta = 0.5$) is used to keep the number of spurious detections low. Then, assuming constant tempo, the rhythmic structure of the input audio signal is estimated by calculating the distribution of inter-onset-intervals. Figure 2 shows the probability density estimate of the inter-onset-intervals from an example file. As can be seen, onset times are highly correlated due to the periodic nature of music events, with clear peaks present around 0.21 s and 0.42 s. We choose the most significant inter-onset intervals $\{T_1, ..., T_K\}$ from this distribution estimate by peak-picking the histogram. This rhythmic information is used in the next section as an additional constraint in our fusion approach in order to obtain a better onset detection performance. This flow of information defines a process where the rhythm is first estimated (bottom-up) and then used in the subsequent onset detection (top-down).

2.4. Fusion with Rhythmic Structure

The goal of our information fusion algorithm is to find the best-scoring set of peaks that are close in time as well as reflecting the rhythmic structure learned from the input signal. The rhythmic structure information $\{T_1, ..., T_K\}$ that relates onsets is added as an additional set of goals to the nonlinear program defined in section 2.2. The resulting cost function is:

$$\begin{aligned} \tilde{C}(\{t_i\}) \quad = \quad & \sum_{i=1}^{N} s_i + \sum_{i=2}^{N} P(t_i - t_{i-1}, T_{\text{ref}}) + \\ & \sum_{i=1}^{N} M(t_i, T_1, ..., T_K) \end{aligned} \tag{4}$$

where $M(t, T_1, ..., T_K)$ is a function that favours peaks that are supported by the estimated rhythmic structure $\{T_1, ..., T_K\}$. If the objective fusion function $\tilde{C}$ has a large value at time $t_i - T_k$ then there is more likely to be an onset at time t_i due to the inter-onset-interval distribution observed in the music signal. We define the gain function $M(t, T_1, ..., T_k)$ as,

$$M(t, T_1, ..., T_k) = \begin{cases} c_{\max}(t) & \text{if } c_{\max}(t) > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $c_{\max}(t)$ represents the maximum fusion scores at time $t - T_k$ with $k = 1, ..., K$,

$$c_{\max}(t) = \max\{C(t - T_1), ..., C(t - T_K)\}. \tag{6}$$

As in section 2.2 we extract the onsets by finding the local maxima of $\tilde{C}(\{t_i\})$. We investigate the benefit of including this rhythmic structure knowledge in Section 4.

3. DATASET AND EVALUATION

A common approach in the evaluation of onset detection algorithms is to use hand-labelled datasets. However, manual annotations introduce ambiguity in the ground truth that makes the analysis and evaluation of the algorithm difficult. The more complex an audio signal, the larger the uncertainty associated with the manual annotations will be. An effective way to obtain a more robust dataset would be to have multiple listeners label each file. We could then compare the performance of our automatic onset detection algorithm with the performance of the mean and best annotator. However, the process of annotating a whole dataset with multiple listeners is very time consuming and often impractical for large datasets.

In order to evaluate our onset fusion algorithm without the additional difficulty of dealing with the uncertainty of manual annotations an onset database generated from MIDI has been created which we make publicly available [2]. The dataset consists of 142 seconds of audio with 482 onsets. The audio is a complex mixture of multiple instruments with no singing and the ground truth onsets are directly extracted from the MIDI files. The size of the dataset is similar to the complex-mixture class in [1] and the dataset recently used in [5].

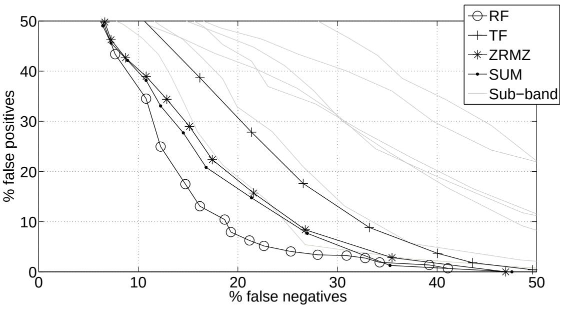

Fig. 3. Comparison of onset detection algorithms: the reference state of the art approach (ZRMZ), the sum of sub-band detection functions (SUM), the individual sub-band onset detection functions (Sub-band) and the temporal (TF) and rhythmic (RF) fusion approaches.

For the evaluation and comparison of onset detection algorithms three measures are usually considered: precision, p, recall, r, and F-measure, f. These evaluation measures are defined as [8]:

$$p \quad = \quad \frac{n_{\text{cd}}}{n_{\text{cd}} + n_{\text{fp}}} \tag{7}$$

$$r \quad = \quad \frac{n_{\text{cd}}}{n_{\text{cd}} + n_{\text{fn}}} \tag{8}$$

$$f \quad = \quad \frac{2pr}{p + r} \tag{9}$$

where n_{cd} is the number of correctly detected onsets, n_{fp} is the number of false positives (detection of an onset when no ground truth onset exists) and n_{fn} is the number of false negatives (missed detections). A correct detection is defined as one occurring within a 50 ms tolerance window of each ground truth onset. Since our aim is not try to identify individual notes, we do not penalise merged onsets.

Using the F-measure we compare the performance of four onset detection approaches: a reference state of the art system, the energy-based approach of Zhou et al [4], which we refer to as ZRMZ (from Zhou, Reiss, Mattavelli and Zoia). We then include an alternative method for combining onset detection functions, defined as the temporal sum of the sub-band onset detection, this we label SUM. To these, we compare the performance of our fusion system without and then with rhythmic knowledge, which we label TF (temporal fusion) and RF (rhythmic fusion) respectively.

4. RESULTS AND DISCUSSION

For each onset detection method the relationship between the rate of false positives (spurious detections) and false negatives (missed detections) is presented in Figure 3. To trace out the performance curve the detection threshold δ (used in peak-picking [1]) was varied between 0 and 1. Better performance is indicated by a shift of the curve to the bottom-left corner of the axes which corresponds to a 0 rate of false positives and negatives.

As can be seen in Figure 3, the performance curve of the rhythmic-based fusion approach is below the curves of the other algorithms under comparison. Based on these results we are able to show, over a range of detection thresholds in the peak-picking process, that the use of rhythmic information for fusing onsets can exceed the state of the art approach (ZRMZ) and the sum of sub-bands approach (SUM). It is noteworthy that the increase in performance for the rhythmic fusion is very pronounced compared to the temporal fusion method (TF). This demonstrates that the addition of a rhythmic constraint into the cost function (as defined in

[2] http://www.gts.tsc.uvigo.es/~ndegara/

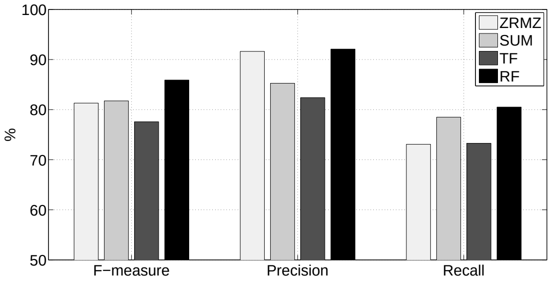

Fig. 4. Maximum F-measure, and Precision and Recall for the detection algorithms under evaluation: the reference state of the art approach (ZRMZ), the sum of sub-band detection functions (SUM) and the temporal (TF) and rhythmic (RF) fusion approaches. The Precision and Recall values are those which correspond to the maximum F-measure.

section 2.4) is crucial when seeking to reduce the number of missed and spurious detections. The temporal fusion algorithm is also the worst performing method. The reason for this is that the number of false positives for this algorithm is very large. We could expect this behaviour since the algorithm processes the whole set of sub-band peak-scores to group only those peaks that are closely related in time. Peak-scores that are not consistently grouped may cause false detections. In our system, the aim of this first temporal fusion step is to provide relevant rhythmic information to be used in the top-down process.

For each applied threshold, a value of F-measure is obtained. Figure 4 presents the maximum F-measure and the values of Precision and Recall corresponding to this F-measure for the rhythmic (RF) and temporal fusion (TF) approaches, the single-band resonator time frequency image (ZRMZ) and the sum of sub-band detection functions (SUM). As we might expect from the performance curve in Figure 3, the best performing algorithm is the rhythmic fusion method. The value of the F-measure of the rhythmic fusion method is 86% which is better than the single-band ZRMZ, 81%, and the sum of sub-band detection functions, 82%. For these values of F-measure, the precision of the ZRMZ and the rhythmic method is 92% in both cases, however the recall of the ZRMZ 73% which is much lower than in the rhythmic case, 81%.

Although we have used audio signals derived from MIDI data to evaluate our algorithm, the accuracy scores obtained are comparable to those based on hand-labelled data [1], [4]. On this basis we believe that our dataset is of sufficient difficulty to reliably test the onset detection algorithms.

In an informal experiment we compared human annotated onsets to the MIDI-derived ground truth on our dataset and discovered a large number of false positives (around 7%) and false negatives (around 15%). As part of our future work, we intend to explore the differences between MIDI-derived ground truth and hand annotated data.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced a framework for onset detection that integrates the information provided by multiple detection functions and rhythmic information relating onsets using a top down-processing approach. We have shown that musical knowledge of periodic structure can successfully be exploited to reduce the number of spurious and missed detections. Results show that the performance is increased when we exploit the rhythmic structure of music signals and that our method is able to outperform the state-of-the-art onset detection algorithm across a wide range of onset detection thresholds. We find the following relative ordering of performance: rhythmic fusion (RF), sum of sub-band onset detection functions (SUM), reference state of the art system (ZRMZ) and then the temporal fusion (TF) approach.

It is important to note that within our framework we do not aim to derive a ‘new’ onset detection function. Instead, our study shows how to combine multiple detection functions exploiting musical knowledge, in particular making use of the inherent rhythmic structure of musical signals. Our approach could use other detection function algorithms as input, such as those in [1] or [9]. Informal experiments have demonstrated that fusion of these features can also lead to improved performance.

In future work, we will explore how to automatically weight or select the individual sub-band detection functions for other music information retrieval tasks such as beat tracking. We also intend to explore extensions to our fusion approach, in particular how to contend with signals exhibiting tempo variation. For such types of signal we plan to include a tempo contour into our objective function building upon the use of local periodicity kernels in [5].

6. REFERENCES

[1] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler, “A tutorial on onset detection in music signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 1035–1047, 2005.

[2] S. Dixon, “Evaluation of audio beat tracking system BeatRoot,” *Journal of New Music Research*, vol. 36, no. 1, pp. 39–51, 2007.

[3] A. Robertson and M. D. Plumbley, “B-Keeper: A beat-tracker for live performance,” in *Proceedings of the International Conference on New Interfaces for musical expression (NIME)*, New York, USA, June, 6–9 2007, pp. 234–237.

[4] R. Zhou, M. Mattavelli, and G. Zoia, “Music onset detection based on resonator time frequency image,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1685–1695, 2008.

[5] Peter Grosche and Meinard Müller, “Computing predominant local periodicity information in music recordings,” in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, 2009.

[6] N. Degara-Quintela, A. Pena, and S. Torres-Guijarro, “A Comparison of Score-level Fusion Rules for Onset Detection in Music Signals,” in *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR’09)*, Kobe, Japan, October 2009.

[7] Daniel P. W. Ellis, “Beat tracking by dynamic programming,” *Journal of New Music Research*, vol. 36, pp. 51–60, 2007.

[8] Simon Dixon, “Onset detection revisited,” in *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06)*, Montreal, Quebec, Canada, Sept. 18–20, 2006, pp. 133–137.

[9] A. Lacoste and D. Eck, “A supervised classification algorithm for note onset detection,” *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 153–153, 2007.