A COMPARISON OF SCORE-LEVEL FUSION RULES FOR ONSET DETECTION IN MUSIC SIGNALS

Norberto Degara-Quintela, Antonio Pena
Department of Signal Theory and Communications
University of Vigo, Spain
{ndegara, apena}@gts.tsc.uvigo.es

Soledad Torres-Guijarro
Laboratorio Oficial de Metroloxía
de Galicia (LOMG), Spain
storres@lomg.net

ABSTRACT

Finding automatically the starting time of audio events is a difficult process. A promising approach for onset detection lies in the combination of multiple algorithms. The goal of this paper is to compare score-level fusion rules that combine signal processing algorithms in a problem of automatic detection of onsets. Previous approaches usually combine detection functions by adding these functions in the time domain. The combination methods explored in this work fuse, at score-level, the peak score information (peak time and onset probability) in order to obtain a better estimate of the probability of having an onset given the probability estimates of multiple experts. Three state-of-the-art spectral-based onset detection functions are used: a spectral flux detection function, a weighted phase deviation function, and a complex domain detection function. Both untrained and trained fusion rules will be compared using a standard data set of music excerpts.

1. INTRODUCTION

The automatic detection of onsets is essential in many applications, including a number of important music information retrieval (MIR) tasks. Onset detection is useful in the analysis of the temporal structure of music as, for example, beat tracking and tempo induction, but it is also important in other relevant tasks such as melody, bass-line and chord extraction.

Finding automatically the starting time of audio events is a difficult process and many onset detection methods exist [1–3]. However, the performance of current detection methods is highly dependent on the nature of the signal as shown in [1]. The reason is that onset detection techniques assume an implicit nature or probability model for the signal to be analyzed. Actually, several well known algorithms can be described in terms of an implicit probability model of the signal [4].

For this reason, it is not expected that a single method will perform accurately for strongly nonstationary signals and audio signals are intrinsically variable in nature. Instead of designing a very complex algorithm, a promising development lies in the combination of multiple methods [5]. In fact, this is most likely the way human perception seems to work [6], using different processing principles for the same purpose so when one of them fails perhaps another succeeds.

Methods that combine time-domain onset detection functions to provide with a more accurate detection have been proposed. However, most of the existing combination schemes use ad-hoc approaches that, for example, choose a particular detection function between two different functions based on the type of onset [7] or a quality measure [8].

Recently, onset detection systems based on machine learning algorithms have been developed. In [9] two Gaussian Mixture Models are used to merge multiple audio features, but the combination of the individual detection functions is still done by a linear weighted sum of the time domain functions. Other approaches merge the detection functions using a time-delay neural network [10, 11].

The integration of tools and information is one of the significant challenges for the field of MIR as discussed in [12] and fusion methods can potentially be used for this purpose. Fusion is an important research area that studies the combination of multiple sources of knowledge to obtain more reliable information [13, 14].

This paper emphasizes the use of information fusion methods to gather the efforts of MIR community which develops multiple signal processing algorithms for the same purpose. In particular, we compare the use of untrained and trained fusion rules to combine, at score-level, the peak information obtained from three spectral-based onset detection functions. Scores represent the estimated time instant and the probability of having an onset at that instant. Hence, our multiple-expert approach aims to calculate a better estimate of that probability given the probability estimates (scores) of multiple experts, which is radically different to adding time-detection functions as previous approaches do. This study is the first work, to our knowledge, that focuses just on the combination of techniques by introducing score-level fusion for onset detection, opening a novel direction to address the problem of combining detection algorithms.

Section 2 introduces the fusion approach to onset detection, describing the structure of the system and the detection functions extracted. Section 3 describes the dataset

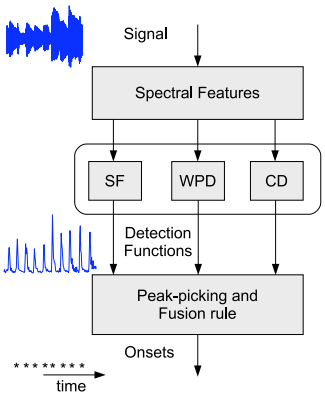

Figure 1. The Multiple-expert paradigm. The system fuses the peak information extracted from three detection functions: the spectral flux measure (SF), the weighted phase deviation (WPD) and the complex domain method (CD).

and the evaluation measures used in the present work. Results are presented in Section 4. And finally, Section 5 contains the conclusions and some ideas for future work.

2. FUSION FOR ONSET DETECTION

Fusion is an important and widely studied area that focuses on the issue of how to combine information to achieve an improved performance. This multiple expert paradigm is based on the combination of various diagnosis to exploit the expertise of the different experts. Score-level fusion combines the different opinions (probability estimates) of the experts to obtain a better estimate of the appropriate a posteriori probability.

Figure 1 shows the multiple expert fusion system that combines the peak information obtained from three state-of-the-art onset detection algorithms. First, the spectrum of the audio signal is calculated using the Short Time Fourier Transform (STFT). Then, three experts derive the detection functions using features extracted from the STFT. Finally, the system combines the peak information obtained from the detection functions using a fusion rule.

2.1 Onset Detection Functions

The detection functions used for fusion in this work are the following spectral-based reduction methods: the spectral flux measure (SF), the weighted phase deviation (WPD) and the complex domain method (CD) described in [2].

All these methods are based on a STFT scheme that applies a Hamming window $h(n)$. Given an audio signal $x(n)$ sampled at $f_s = 44.1$ kHz, the kth frequency bin of the nth spectrum frame $X(n,k)$ is given by:

$$X(n,k) = \sum_{m=-\frac{N}{2}}^{m=\frac{N}{2}-1} x(nh+m)h(m)\exp^{-\frac{j2\pi km}{N}} \quad (1)$$

In our experiments, the window size in samples is $N = 2048$ (46 ms) and the hop size $h = 441$ (10 ms).

The spectral flux (SF) measures the distance between successive short-time Fourier spectra:

$$SF(n) = \sum_{m=-\frac{N}{2}}^{m=\frac{N}{2}-1} H(|X(n,k)| - |X(n-1,k)|) \quad (2)$$

where $H(x) = \frac{x+|x|}{2}$ is a half-wave rectifier. This function is used to emphasize onsets rather than offsets since the sum is restricted to those frequencies where the spectral difference is positive and an increase of energy exists.

In order to add phase information in this system of multiple experts, the weighted phase deviation reduction method has also been considered. The rate of change of phase is an estimation of the instantaneous frequency and abrupt changes in the instantaneous frequency may suggest a potential onset. The weighted phase deviation (WPD) reduction method takes the mean of the absolute value of the instantaneous frequency difference weighted by the magnitude of the spectra:

$$WPD(n) = \frac{1}{N}\sum_{m=-\frac{N}{2}}^{m=\frac{N}{2}-1} |X(n,k)||\varphi''(n,k)| \quad (3)$$

where $\varphi''(n,k)$ is the second derivative of the 2π-unwrapped phase of the Fourier spectra $X(n,k)$.

Finally, the complex domain detection function considers jointly both magnitude and phase to search for transients on the signal. The spectral component $X(n,k)$ can be predicted from the previous frame spectra magnitude and phase change:

$$\widehat{X}(n,k) = |X(n-1,k)|e^{\varphi(n-1,k)+\varphi'(n-1,k)} \quad (4)$$

The complex domain (CD) detection function is defined as the sum of the absolute deviations from the predicted spectral values $\widehat{X}(n,k)$,

$$CD(n) = \sum_{m=-\frac{N}{2}}^{m=\frac{N}{2}-1} |X(n,k) - \widehat{X}(n,k)| \quad (5)$$

Normalization is a key step in fusion, therefore each of the detection functions is normalized to have a mean 0 and standard deviation of 1.

2.2 The Multiple-expert Architecture

In this approach, where multiple algorithms are combined to accomplish the same goal and can potentially interact to adapt its behavior, the architecture is very important. In this sense, blackboard modeling, an approach taken from artificial intelligent systems, has been successfully applied to other relevant applications such as computational auditory scene analysis [15] and polyphonic music transcription [16]. In a blackboard model, experts communicate using a common database what allows to pursue multiple lines of analysis at the same time and to adapt the strategies to a particular problem context.

The multiple-expert approach described in this paper has been developed within a blackboard-agent framework. Although the number of experts used in this paper is small, the blackboard-agent framework will probably be useful when combining many more experts, by implementing top-down processing where results coming from fusion are fed back into experts to improve individual results.

2.3 Peak Selection

Peaks are selected from the onset detection functions by peak-picking the local maxima. We apply the peak-picking algorithm used in [2] to obtain the peak-score information used for fusion: the peak time and the estimated probability of having an onset at that time.

A peak at time $t = \frac{nh}{f_s}$ (where n is the current sample, h the hop length and f_s the sampling frequency) is chosen as a relevant peak if the peak is a local maximum and the detection function is larger than a threshold above the local mean of the detection function $f(n)$, this is:

$$f(n) \geq f(m) \text{ for } m \text{ such that } n - w \leq m \leq n + w \quad (6)$$

$$f(n) - \frac{\sum_{m=n-lw}^{n+w} f(m)}{lw + w + 1} \geq \delta \quad (7)$$

where w is the size of the window used to find local maxima, l is a weighting factor to calculate the mean over a larger range before the peak (emphasizing onsets rather than offsets) and δ is the threshold.

Peak scores are normalized by subtracting the calculated local mean to the peak value of the detection function as given in equation (7).

The values of the peak-picking parameters have a large impact on the results. Hence, we follow the approach chosen in [1] and [2] selecting the parameters that maximize the F-measure, a performance measure defined in Section 3.

2.4 Fusion

Onsets in the original signal are related to peaks in the detection functions, therefore the normalized peak scores and times pairs are selected by using the mean-filter peak-picking algorithm described above. Peak scores and time stamps from the three experts are grouped in time frames of $50\ ms$ and 50% overlap. If a given expert proposes several peaks within the merging frame, the peak with the highest score is selected.

Let $F(l) = \{f_{sf} f_{pd} f_{cd}\}$ and $T(l) = \{t_{sf} t_{pd} t_{cd}\}$ be, respectively, the peak scores and time stamps for each expert in the grouping time frame l. The proposed system fuses this peak information using the rules described below and classifies the frame as an onset or non-onset frame. The parameters of the fusion algorithms are chosen so as to maximize the performance of the fusion system.

Voting is perhaps one of the oldest strategies for decision making. The voting mechanism counts the number of expert scores that are higher than a given threshold and a consensus pattern is applied. A grouping frame can be classified as an onset-frame if any, the majority or all (unanimity) the experts exceed the threshold.

The sum rule simply adds the normalized expert scores in the grouping frame to obtain a better estimate of the a posteriori onset probability. A frame is labeled as an onset-frame if the resulting sum score exceeds a threshold.

Trained fusion strategies are also explored in this paper. In particular, we evaluate the performance of a K-Nearest Neighbor (K-NN) rule and a Support Vector Machine (SVM) with RBF kernel using cross-validation. The parameters of the RBF kernel are selected using a grid-search technique.

Grouping peak information in overlapping time frames generates doubled detections, therefore the output of the fusion rule is post processed to remove doubled onsets.

3. DATASET AND EVALUATION METHODOLOGY

The evaluation of the proposed fusion approaches is performed using the annotated dataset used in [1]. The dataset is composed of excerpts of different musical styles classified into the following categories: pitched non-percussive (PN), pitched percussive (PP), non-pitched percussive (NP) and complex mixtures (CM). This allows to test the algorithms on different classes of audio signals. There is a total of 1060 onsets.

The majority of the literature reporting results on onset detection shows a lack of proper statistical evaluation. Few works report standard deviations to give an idea of the variability of the results and most of them rely on mean performances only. Fortunately, a proper statistical hypothesis testing methodology has been adopted in MIREX 2008.

Hence, we decided to segment the original signals into homogeneous folds to evaluate the accuracy of our system using K-fold cross-validation. Cross-validation allows the statistical evaluation of the performance measures, enabling the estimation of confidence intervals [17]. We used different cross-validation files for each category, with no overlap between folds. The number of folds were 14 (CM), 12 (NP), 12 (PN) and 14 (PP).

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:

$$P = \frac{n_{cd}}{n_{cd} + n_{fp}} \quad (8)$$

$$R = \frac{n_{cd}}{n_{cd} + n_{fn}} \quad (9)$$

$$F = \frac{2PR}{P + R} \quad (10)$$

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). Due to the reliability of hand-labeled annotations, a time tolerance of 50 ms is usually assumed. This means that an onset is considered to be correctly matched if the detected onset is within

50 ms of the ground truth onset time. In addition, we do not penalized merged onsets since we do not try to identify individual notes.

As discussed in Section 2.3, peak-picking and fusion rule parameters are chosen so as to maximize the F-measure, which assigns the same significance to false positives and false negatives.

4. RESULTS AND DISCUSSION

Table 1 compares the results of the best individual experts and the proposed fusion rules on the different datasets. We choose the best expert for comparison because fusion always performed better than the worst expert in our experiments. In addition, we want to show that fusion can obtain even better results than the best expert and that fusion performance is not limited by the worst expert.

Total performance do not show enough information to compare different approaches and a proper statistical analysis is essential to fully understand how the different methods perform. Hence, Table 1 shows mean values and the 95% confidence interval for the F-measure using cross-validation.

As it can be seen in this table, fusion rules are able to achieve better performance than the best of the experts. For the PN and PP datasets, the relative increase in F-measure is important considering that the performance of the best of the experts is already high. Hence, the accuracy of the fusion algorithms is not limited by the worst of the experts and fusion achieves an improvement in performance by exploiting consensus diagnosis of the three experts.

For the NP and CM cases, the increase in performance given by the fusion rules is not significant. In fact, the performance is limited by the number of false negatives because there is a number of onsets that are not detected for any of the the experts. To exploit the benefits of fusion, experts should be as diverse as possible meaning that onset detection functions should be accurate and should not make coincident errors.

It is noteworthy that fusion has reduced the F-measure deviation in the PN and PP datasets but is still large for the NP and CM datasets. A large deviation means that fusion obtains good results for some of the folds but the performance is very low for other folds. In this sense, the performance could potentially be increased if we were able to identify the quality of the detection functions and apply different fusion strategies based on this quality measure.

We turn now to discuss the different approaches for fusion. Simple fusion rules obtain better results than trained fusion rules. The size of the test sets is small and both the K-NN and SVM approaches suffer from overfitting. In addition, the SVM achieves better performance than the K-NN except for the CM case. Finally, the SVM achieves very good results for the PP case, probably because the number of samples required to learn the task of identifying PP onsets is low.

We followed the statistical evaluation methodology proposed in [17] and we assumed a t-distribution for the sample mean estimator of the F-measure (the number of folds for cross-validation was less than 30). However, performance depends on various factors such as the set size, composition and the choice of the samples. Another interesting accuracy measure would be the Weighted Error Rate (WER), widely used in biometrics. In this case, a specific method for the calculation of confidence intervals for the total WER, not the mean, is already defined in [18]. This method reduces the performance dependency of these factors. The WER, a error measure widely used in biometrics, is defined as:

$$WER(R) = \frac{f_n + Rf_p}{1 + R} \tag{11}$$

where f_n and f_p are the false negatives and positives rates. The parameter R allows to balance the significance of the false positives and false negatives in the error measure which could be of interest in some applications and useful to compare algorithms at different operating points. Therefore, the WER can be an appropriate measure for the statistical evaluation of music information retrieval experiments.

5. CONCLUSIONS AND FUTURE WORK

The originality of this contribution is the introduction of score-level fusion strategies for onset detection, looking at the problem of combining onset information as a multiple-expert fusion problem. Our approach aims to calculate a better estimate of that probability given the probability estimates of multiple experts, which is radically different to adding time-detection functions as previous approaches do. This study is the first work, to our knowledge, that focuses just on the combination of techniques by introducing score-level fusion for onset detection, opening a novel direction to address the problem of combining detection algorithms.

This paper compares untrained and trained fusion rules on four sets of different music styles. Results show how information fusion rules can lead to a higher performance when combining multiple signal processing algorithms designed for onset detection. However, the increase in performance seems to be not important if experts are not diverse. Simple fusion rules show better performance than trained rules due to, probably, the small number of samples available for training.

In addition, a performance measure widely used in biometrics has been proposed. The Weighted Error Rate allows to balance the significance of the false positives and false negatives in the error measure and a specific method for the calculation of the confidence intervals of the total error rate is already defined.

In future work, we will include more experts to exploit diversity in the information fusion process. In addition, the cross-validation analysis showed a high deviation of the F-measure for complex signals. This means that the performance of the experts is highly dependent on the conditions of the signal. To face this problem, we will explore quality-based information fusion which basically weights scores according to the quality of the expert's detection functions.

	PN data			PP data			NP data			CM data		
	P	R	**F**	P	R	**F**	P	R	**F**	P	R	**F**
B.E.	93.8	98.1	$\mathbf{95.7 \pm 5.1}$	97.4	98.5	$\mathbf{97.8 \pm 1.7}$	99.5	94.5	$\mathbf{96.7 \pm 5.5}$	89.4	89.6	$\mathbf{88.8 \pm 6.7}$
Vot.	99.1	95.6	$\mathbf{97.3 \pm 2.4}$	98.4	98.8	$\mathbf{98.6 \pm 1.0}$	96.9	96.7	$\mathbf{96.7 \pm 5.7}$	91.0	88.5	$\mathbf{89.2 \pm 7.5}$
Sum	99.1	95.6	$\mathbf{97.3 \pm 2.4}$	99.8	98.6	$\mathbf{99.2 \pm 0.9}$	98.0	94.6	$\mathbf{96.2 \pm 5.5}$	93.9	85.4	$\mathbf{88.9 \pm 7.0}$
KNN	91.4	96.4	$\mathbf{93.5 \pm 4.3}$	95.7	98.0	$\mathbf{96.7 \pm 1.5}$	94.2	92.6	$\mathbf{93.2 \pm 8.0}$	88.2	82.9	$\mathbf{84.3 \pm 10}.4$
SVM	92.2	98.1	$\mathbf{94.7 \pm 5.6}$	99.5	98.5	$\mathbf{99.0 \pm 1.0}$	96.8	95.6	$\mathbf{96.2 \pm 6.3}$	84.0	84.8	$\mathbf{83.5 \pm 8.9}$

Table 1. Performance comparison of the score-fusion rules and the best individual expert (B.E.), showing precision (P), recall (R) and F-measure (F), for the different data sets. The table shows the mean and 95% confidence interval for the F-measure using K-fold cross-validation.

We will also intend to use a larger dataset to avoid overfiting in trained fusion rules. Finally, we will consider information fusion in other relevant problems such as beat tracking and tempo induction.

6. ACKNOWLEDGMENTS

Thanks to Juan Pablo Bello for kindly providing the dataset. This work has been partially supported by the Spanish MEC, ref. TEC2006-13883-C04-02, under the project AnClaS3 "Sound source separation for acoustic measurements".

7. 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," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 1035–1047, 2005.

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

[3] N. Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *In AES Convention 118*, no. 6363, 2005.

[4] S. Abdallah and M. Plumbley, "Probability as metadata: Event detection in music using ICA as a conditional density model," in *ica03*, Nara, Japan, Apr. 2003, pp. 233–238.

[5] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1832–1844, 2006.

[6] A. Bregman, "Psychological data and computational asa," in *Computational auditory scene analysis*, D. Rosenthal and H. Okuno, Eds. Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc., 1998.

[7] 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.

[8] Y. S. Wan-Chi Lee and C.-C. J. Kuo, "Musical onset detection with linear prediction and joint features," in *2007 MIREX contest results*, 2007.

[9] C.-C. Toh, B. Zhang, and Y. Wang, "Multiple-Feature Fusion Based Onset Detection for Solo Singing Voice," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, Philadelphia, USA, September 2008.

[10] 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.

[11] N. Degara-Quintela, A. Pena, M. Sobreira-Seoane, and S. Torres-Guijarro, "A mixture-of-experts approach for note onset detection," in *126th AES Convention*, Munich, Germany, May 2009.

[12] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[13] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, July 2004.

[14] J. Kittler, "Combining classifiers: A theoretical framework," *Pattern Analysis & Applications*, vol. 1, no. 1, pp. 18–27, March 1998.

[15] D. P. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, MIT Department of Electrical Engineering and Computer Science, 1996, i.

[16] J. P. Bello, "Towards the automated analysis of simple polyphonic music: A knowledge-based approach," Ph.D. dissertation, King's College London - Queen Mary, University of London, 2003, i.

[17] A. Flexer, "Statistical evaluation of music information retrieval experiments," *Journal of New Music Research*, vol. 35, no. 2, pp. 113–120, June 2006.

[18] N. Poh and S. Bengio, "Estimating the confidence interval of expected performance curve in biometric authentication using joint bootstrap," Tech. Rep., 2006.