ONSET DETECTION BASED ON FUSION OF SIMPLS AND SUPERFLUX

Zhengchen Zhang[1], Dong-yan Huang[1], Renbo Zhao[2] and Minghui Dong[1]
[1]Human Language Technology Department,
Institute for Infocomm Research (I2R), A*STAR, Singapore
[2]Department of Electrical and Computer Engineering,
National University of Singapore

ABSTRACT

In this paper, an onset detection method based on Simple Partial Least Squares (SIMPLS) is proposed and a system of fusing SIMPLS and SuperFlux is introduced. SIMPLS is an efficient approach to partial least squares regression which has been applied to classification tasks. To detect onsets in an audio file, the file is sampled discretely into frames and the signals are transferred into frequency domain. The information is then fed into SIMPLS to predict a short time frame is onset or not. A score is obtained for each frame as the possibility of being an onset, and post processing is conducted to pick the peaks. SuperFlux is a robust and fast onset detection method proposed by S. Böck et al. We combine the detection results of SIMPLS and SuperFlux on the decision level. Experimental results demonstrate that the proposed method could generate better F_1-measure value than the individual methods.

1. INTRODUCTION

Onset detection aims to detect the starting instant of an event in an audio signal [8]. Many methods have been proposed in the field, and they can be divided into two classes: supervised learning and unsupervised learning methods. As there will be a rise in energy in the onset attack phase, one can detect the energy increase by calculating the differences between adjacent audio frames [2, 3, 8]. In [8], three different kinds of information: phase, magnitude, and pitch were used to detect the onsets of pitched instruments. In [2,3], unsupervised methods with online detection capabilities were discussed. With unsupervised learning methods, useful features can be selected by experts to detect onsets in some specific types of music. It is difficult to employ more features to make the methods more widely applicable. Machine learning methods can solve this problem by combining different features. Neutral networks and Hidden Markov Models have been applied to onset detection in [1, 5, 6]. However, it will take a long time to train a model if the dimension of the feature matrix is huge. In this paper, we introduce a fast onset detection method based on SIMPLS. Also, a system fusing both supervised and unsupervised learning methods is proposed to improve the detection accuracy.

2. SYSTEM DESCRIPTION

In this section, we will first describe an onset detection method based on SIMPLS. The fusion system combining SIMPLS and SuperFlux is then introduced.

2.1 Onset detection based on SIMPLS

To detect onsets in an audio, one can segment the file into short time frames, and classify each frame is onset or not. It is an imbalanced classification problem as there are very few onset frames comparing to the number of frames that are not onsets. Imbalanced classification problems will lead to failure of some standard methods [7]. SIMPLS has been proved to be efficient in solving the problem, and it is faster than many existing classification methods [9]. Here we employ SIMPLS to detect onsets in an audio file.

2.1.1 Feature extraction

An input audio file is segmented into short overlapping frames with window size 512, 1024, 2048, 3072 and 4096 respectively. The frame rate is set to be 100 frames per second. A Hann window with same length is used to weight the frames. Then the frames are transformed into the frequency domain with the Discrete Fourier Transform (DFT). Filtering is applied with 24 bands per octave. A logarithmic representation is chosen to transform the feature values to a range more suitable for training the SIMPLS model. The differences between adjacent frames are also included in the features as they have been proved to be very useful in onset detection [1, 3]. The features are normalized before they are used to train a SIMPLS model by transforming every column of the feature to a random variable with zero mean and unit variance using

$$\widetilde{x}_{i,j} = (x_{i,j} - \mu_j)/\delta j \quad (1)$$

where $x_{i,j}$ is an item of the feature matrix. The mean and variance of items in jth column are represented by μ_j and δ_j.

To train a SIMPLS model, we need to prepare the label of each frame. Given the labelled onset times T of an audio, we set the label of a frame to 1 if the time of a frame

t_f satisfies that $\text{abs}(t_f - t_i) \leq 0.025s$ where $t_i \in T$. Otherwise, the label is set to be -1.

2.1.2 SIMPLS training and testing

Partial Least Squares (PLS) is a dimension reduction method that has been adapted for high dimensional classification problems [9]. SIMPLS is an efficient approach to PLS due to the avoidance of matrix inverse calculation [4]. Given a feature matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$ and a label matrix $\mathbf{y} \in \mathbb{R}^{N \times 1}$, SIMPLS aims to find a linear projection [4]

$$\widehat{\mathbf{Y}} = \mathbf{XB} \tag{2}$$

where N is the number of samples and M is the dimension of features. Here $\mathbf{y}$ is one dimensional as the onset detection is a binary classification problem. The solution is obtained by extracting the successive orthogonal factors of $\mathbf{X}$ and $\mathbf{Y}$,

$$\mathbf{t}_a = \mathbf{X}_0\mathbf{r}_a \tag{3}$$

and

$$\mathbf{u}_a = \mathbf{Y}_0\mathbf{q}_a, a = 1, 2, ..., A \tag{4}$$

with the following four restrictions, where $\mathbf{X}_0 = \mathbf{X} - \text{mean}(\mathbf{X})$, $\mathbf{Y}_0 = \mathbf{Y} - \text{mean}(\mathbf{Y})$, and $A \leq M$.

1. The covariance of $\mathbf{u}_a$ and $\mathbf{t}_a$ is maximized: $\max(\mathbf{u}_a'\mathbf{t}_a) = \max(\mathbf{q}_a'(\mathbf{Y}_0'\mathbf{X}_0)\mathbf{r}_a)$;
2. $\mathbf{r}_a$ is normalized: $\mathbf{r}_a'\mathbf{r}_a = 1$;
3. $\mathbf{q}_a$ is normalized: $\mathbf{q}_a'\mathbf{q}_a = 1$; and
4. The scores are orthogonal to each other: $\mathbf{t}_b'\mathbf{t}_a = 0$ for $a > b$.

The algorithm of extracting scores and loadings of $\mathbf{X}$ and $\mathbf{Y}$ is shown in Algorithm 1 [4, 9, 11].

In the algorithm, the score vector $\mathbf{t}_a$ is normalized by $\mathbf{t}_a = \mathbf{t}_a/\sqrt{\mathbf{t}_a' * \mathbf{t}_a}$. According to the restriction (4), we have $\mathbf{T}'\mathbf{T} = \mathbf{I}$ where $\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_A]$. To predict the label of the samples,

$$\widehat{\mathbf{Y}}_0 = \mathbf{TT}'\mathbf{Y}_0 = \mathbf{X}_0\mathbf{RR}'\mathbf{X}_0'\mathbf{Y}_0 = \mathbf{X}_0\mathbf{RR}'\mathbf{S}_0 \tag{5}$$

Hence, $\mathbf{B}$ in (2) can be written as:

$$\mathbf{B} = \mathbf{R}(\mathbf{R}'\mathbf{S}_0) = \mathbf{R}(\mathbf{T}'\mathbf{Y}_0) = \mathbf{RQ}' \tag{6}$$

The labels of new samples can be predicted by

$$\widehat{\mathbf{Y}}^* = \mathbf{X}_0^*\mathbf{B} \tag{7}$$

where $\mathbf{X}_0^* = \mathbf{X}^* - \text{mean}(\mathbf{X}^*)$ and $\mathbf{X}^*$ is the new feature matrix.

The prediction of a new sample can be written in another format. PLS will find the score and loading vectors in (3) and (4). The inner relation between $\mathbf{X}$ and $\mathbf{Y}$ can be estimated by the regression coefficient $\mathbf{b}$ via the latent variables [4]:

$$\widehat{\mathbf{u}}_a = \mathbf{b}_a\mathbf{t}_a \tag{8}$$

$$\mathbf{b}_a = \mathbf{u}_a'\mathbf{t}_a/(\mathbf{t}_a'\mathbf{t}_a) \tag{9}$$

Algorithm 1 SIMPLS Training

Input: Feature set $\mathbf{X}$, Label $\mathbf{y}$, and Number of components A
Variables: Projection matrix $\mathbf{R}$, score vectors $\mathbf{T}$ and $\mathbf{U}$, loading $\mathbf{P}$ and $\mathbf{Q}$
$\mathbf{R} = []; \mathbf{V} = []; \mathbf{Q} = []; \mathbf{T} = []; \mathbf{U} = [];$
$\mathbf{y} = [y_1, y_2, ..., y_N]'; \mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]'$
$\mathbf{y}_0 = \mathbf{y} - \text{mean}(\mathbf{y}); \mathbf{X}_0 = \mathbf{X} - \text{mean}(\mathbf{X});$
$\mathbf{S} = \mathbf{X}_0'\mathbf{y}_0$
for $i = 1$ to A **do**
  $\mathbf{q}_i =$ dominant eigenvectors of $\mathbf{S}'\mathbf{S}$
  $\mathbf{r}_i = \mathbf{S} * \mathbf{q}_i$
  $\mathbf{t}_i = \mathbf{X}_0 * \mathbf{r}_i$
  $\text{normt}_i = \text{SQRT}(\mathbf{t}_i'\mathbf{t}_i)$
  $\mathbf{t}_i = \mathbf{t}_i/\text{normt}_i$
  $\mathbf{r}_i = \mathbf{r}_i/\text{normt}_i$
  $\mathbf{p}_i = \mathbf{X}_0' * \mathbf{t}_i$
  $\mathbf{q}_i = \mathbf{y}_0^T * \mathbf{t}_i$
  $\mathbf{u}_i = \mathbf{y}_0 * \mathbf{q}_i$
  $\mathbf{v}_i = \mathbf{p}_i$
  if $i > 1$ **then**
    $\mathbf{v}_i = \mathbf{v}_i - \mathbf{V} * (\mathbf{V}' * \mathbf{p}_i)$
    $\mathbf{u}_i = \mathbf{u}_i - \mathbf{T} * (\mathbf{T}' * \mathbf{u}_i)$
  end if
  $\mathbf{v}_i = \mathbf{v}_i/\mathbf{SQRT}(\mathbf{v}_i' * \mathbf{v}_i)$
  $\mathbf{S} = \mathbf{S} - \mathbf{v}_i * (\mathbf{v}_i' * \mathbf{S})$
  $\mathbf{r}_i, \mathbf{t}_i, \mathbf{p}_i, \mathbf{q}_i, \mathbf{u}_i,$ and $\mathbf{v}_i$ into
  $\mathbf{R}, \mathbf{T}, \mathbf{P}, \mathbf{Q}, \mathbf{U},$ and $\mathbf{V}$, respectively.
end for
$\mathbf{B} = \mathbf{R} * \mathbf{Q}'$

Algorithm 2 Predicting labels for new samples using PLS.

Input: New feature matrix $\mathbf{X}$; projection matrix $\mathbf{R}$, regression coefficients $\mathbf{b}$, loading $\mathbf{P}$ and $\mathbf{Q}$ obtained by Algorithm 1 on the training set.
Output: Predicted Label $\hat{Y}$ of $\mathbf{X}$

$\mathbf{X}_0 = \mathbf{X} - \text{mean}(\mathbf{X})$
for $i = 1$ to A **do**
  $\mathbf{t}_i = \mathbf{X}_{i-1}\mathbf{r}_i;$
  $\mathbf{X}_i = \mathbf{X}_{i-1} - \mathbf{t}_i\mathbf{p}_i';$
end for
$\hat{\mathbf{Y}} = \text{sign}(\sum_{i=1}^{A} b_i\mathbf{t}_i\mathbf{q}_i')$

Hence, the algorithm of predicting labels for new samples is summarized in Algorithm 2.

From Algorithm 2, we have

$$\hat{\mathbf{Y}} = \text{sign}(\sum_{i=1}^{A} \mathbf{b}_i \mathbf{t}_i \mathbf{q}_i') = \text{sign}(\sum_{i=1}^{A} \mathbf{m}_i \mathbf{t}_i) = \text{sign}(\mathbf{m} \cdot \mathbf{t}) \tag{10}$$

One can see that the label $\widehat{\mathbf{Y}}$ is a function of the score vectors $\mathbf{t}$.

2.1.3 Post-processing

As described above, a SIMPLS model generates a score for each frame. Instead of using hard decision $\text{sign}(\mathbf{m} \cdot \mathbf{t})$ in (10), we use the value $\mathbf{m} \cdot \mathbf{t}$ directly as the possibilities of being an onset of a frame. The scores are normalized by (1), and then the post-processing method in [3] is used to detect peaks. A frame f is selected as an onset if the score s_f fulfills the following conditions:

$$s_f = \max(s_f(f - n_1 : f + n_2)) \tag{11}$$

$$s_f \geq \text{mean}(s_{f-n_3:f+n_4}) + threshold \tag{12}$$

$$f - f_{last\_onset} > n_5 \tag{13}$$

where $n_i, i = 1, ..., 5$ are tunable peak-picking parameters.

2.2 Fusion of SIMPLS and SuperFlux

SuperFlux [3] is a state-of-the-art onset detection method which calculates the difference between two near short-time spectra and is optimized for music with much vibrato. For each audio file, SIMPLS and SuperFlux detect the onsets separately and then the results are fused on the decision level. We combine the time points detected by SIMPLS and SuperFlux with Algorithm 3. If two time points are near to each other, we calculate the average value of the time and put it into the final result. Otherwise, we put the time point into the final result directly.

Algorithm 3 Fusion of SIMPLS and SuperFlux.

Input: Onset time list T_{sp} and T_{sf} obtained by SIMPLS and SuperFlux respectively; A threshold θ
Output: The final decision T
index1 = 0
index2 = 0
T=[]
while index1<len(T_{sp}) and index2<len(T_{sf}) **do**
  t1 = T_{sp}[index1]
  t2 = T_{sf}[index2]
  if abs(t1-t2)< θ **then**
    T.append((t1+t2)/2)
    index1 += 1
    index2 += 1
  else if t1>t2 **then**
    T.append(t2)
    index2 += 1
  else if t1<t2 **then**
    T.append(t1)
    index1 += 1
  end if
end while
if index1<len(T_{sp}) **then**
  for temp in T_{sp}[index1:end] **do**
    T.append(temp)
  end for
  else if index2<len(T_{sf})
  for temp in T_{sf}[index2:end] **do**
    T.append(temp)
  end for
end if

3. EXPERIMENTAL RESULTS

3.1 Datasets

Three data sets are used in this work: Sound Onset Labellizer (SOL), onset detection database (ODB), and NYU. The SOL data set [1] is annotated by the sound onset labellizer proposed in [10]. It has 18 audio files (with labels) and 671 onsets. The ODB data set is downloaded from internet [2] which consists of 19 audio files and 2155 onsets. The sampling rate of files in ODB is 22.05 kHz, and we transferred them into 44.1 kHz using Adobe Audition. The NYU data set is shared by Professor Juan Pablo Bello [3] in New York University (NYU). It has 23 audio files and 1060 onsets. In total, we have 60 audio files as our training and testing data set.

[1] http://www.tsi.telecom-paristech.fr/aao/en/2011/07/13/sound-onset-labellizer/

[2] http://grfia.dlsi.ua.es/cm/projects/prosemus/database.php

[3] https://files.nyu.edu/jb2843/public/Home.html

Table 1. Experimental results of different methods on the data sets with 10-fold cross validation.

Methods	Precision	Recall	F_1-measure
SuperFlux	0.719	0.781	0.749
SIMPLS	0.808	0.690	0.744
Fusion	0.719	0.808	0.761

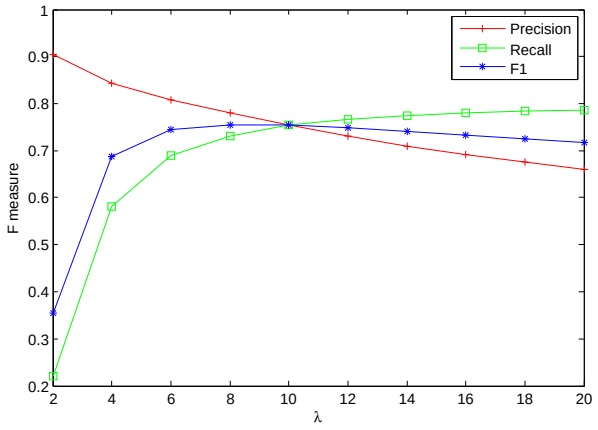

Figure 1. System performance of SIMPLS with different λ.

3.2 Results

In our experiments, an onset is detected correctly if it is within a 50 ms window around the ground truth onset position. We conduct 10-fold cross validation and report the results in Table 1. As SuperFlex is an unsupervised learning method, the results are same for each round of cross validation.

One can see that Fusion could generate 1 percent higher F_1-measure value than SuperFlux. The recall increases after fusing, which indicates that the fusion system could find out more onsets than individual methods as it combines the onsets found by them. The precision is almost same as SuperFlux and lower than SIMPLS. This demonstrates that the fusion did not bring too much false positive items. It is worth noting that the precision of SIMPLS is much higher than its recall. We intended to obtain this result by adjusting the threshold in (12). SIMPLS are expected to capture few but accurate onsets to be a supplement of SuperFlux.

By adjusting the threshold, SIMPLS could obtain competitive F_1-measure value with SuperFlux. Experiments are conducted by setting the threshold be $\max(S)/\lambda$ where S is the score set of all the frames. We set $\lambda = 2, 4, 6, ..., 20$. The results are shown in Fig. 1. With the increasing of λ, the threshold decreases. The recall value increases as we can find more onsets. However, the precision decreases because more false positive items are detected. When $\lambda = 8$, the F_1 value achieved to the highest 0.755.

4. CONCLUSION

In this paper, a system fusing SIMPLS and SuperFlux for onset detection has been proposed. Experimental results have demonstrated that the fusion system could obtain little improvement than single methods. In this work, only spectra features and their differences are considered in our system. More features will be included in the future work. Both SIMPLS and SuperFlux methods are fast, which is suitable for realtime onset detection. More comparative experiments will be conducted to demonstrate the efficiency of the two methods and the fusion system.

5. REFERENCES

[1] Sebastian Böck, Andreas Arzt, Florian Krebs, and Markus Schedl. Online realtime onset detection with recurrent neural networks. In *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12)*, page 4, 2012.

[2] Sebastian Böck, Florian Krebs, and Markus Schedl. Evaluating the online capabilities of onset detection methods. In *ISMIR*, pages 49–54, 2012.

[3] Sebastian Böck and Gerhard Widmer. Maximum filter vibrato suppression for onset detection. In *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013)*, 2013.

[4] Sijmen de Jong. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993.

[5] Norberto Degara, Matthew EP Davies, Antonio Pena, and Mark D Plumbley. Onset event decoding exploiting the rhythmic structure of polyphonic music. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1228–1239, 2011.

[6] Florian Eyben, Sebastian Böck, Björn Schuller, and Alex Graves. Universal onset detection with bidirectional long short-term memory neural networks. In *ISMIR*, pages 589–594, 2010.

[7] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[8] André Holzapfel, Yannis Stylianou, Ali Cenk Gedik, and Baris Bozkurt. Three dimensions of pitched instrument onset detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1517–1527, 2010.

[9] Dong-Yan Huang, Shuzhi Sam Ge, and Zhengchen Zhang. Speaker state classification based on fusion of asymmetric simpls and support vector machines. In *INTERSPEECH*, pages 3301–3304, 2011.

[10] Pierre Leveau and Laurent Daudet. Methodology and tools for the evaluation of automatic onset detection algorithms in music. In *In Proc. Int. Symp. on Music Information Retrieval*. Citeseer, 2004.

[11] Zhengchen Zhang. *Data Analysis for Emotion Identification in Text*. PhD thesis, National University of Singapore, 2013.