Musical Onset Detection Using Constrained Linear Reconstruction

Che-Yuan Liang, Li Su, *Member, IEEE*, and Yi-Hsuan Yang, *Member, IEEE*

***Abstract*—This letter presents a multi-frame extension of the well-known spectral flux method for unsupervised musical onset detection. Instead of comparing only the spectral content of two frames, the proposed method takes into account a wider temporal context to evaluate the dissimilarity between a given frame and its previous frames. More specifically, the dissimilarity is measured by using the previous frames to obtain a linear reconstruction of the given frame, and then calculating the rectified, l_2-norm reconstruction error. Evaluation on a dataset comprising 2,169 onset events of 12 instruments shows that this simple idea works fairly well. When a non-negativity constraint is imposed in the linear reconstruction, the proposed method can outperform the state-of-the-art unsupervised method SuperFlux by 2.9% in F-score. Moreover, the proposed method is particularly effective for instruments with soft onsets, such as violin, cello, and ney. The proposed method is efficient, easy to implement, and is applicable to scenarios of online onset detection.**

***Index Terms*—Exemplar, linear reconstruction, musical onset.**

I. Introduction

ONSET detection is the process of locating the starting points of musically relevant events in a music signal [1]–[4]. Onset detection is a fundamental task in music information retrieval, with numerous applications such as note segmentation [5], [6], automatic music transcription [7], beat tracking [8], and interactive musical accompaniment [9], amongst others. The main challenge of musical onset detection is to build a robust algorithm that can deal with all types of variability found in music, encompassing different instruments, playing techniques, music styles, and the presence of concurring notes [3]. In consequence, despite that great progress have been made in recent years, musical onset detection remains an active area of academic research [10]–[20].

Many algorithms have been proposed in the literature and evaluated in the Audio Onset Detection task of the annual Music Information Retrieval Evaluation eXchange (MIREX) [21], including unsupervised and supervised methods. Up to date the state-of-the-art methods are based on the recurrent neural networks (RNN) [15]–[17], which is a supervised learning algorithm. Being different from this, the focus of this paper is on unsupervised algorithms, which do not require labeled data. Our specific goal is to develop a new way to compute the *onset detection function* (ODF), a time series that ideally exhibits sharp peaks at onset times, using signal processing techniques. Given the ODF, it is assumed that one can apply *peak picking* methods to identify the onsets [1]–[3].

Among the various kinds of unsupervised onset detection methods, the most popular and widely-investigated ones might be the *spectral flux* (SF) methods, which formulate the ODF as a distance between successive short-time spectral features [1], [2]. For example, Duxbury *et al.* [22] proposed to take the l_2-norm on the *rectified* difference of the magnitude Fourier spectra between the current frame and the previous one, taking into account only the frequencies where there is an increase in energy so as to emphasize onsets rather than offsets. Recently, Böck and Widmer [13] proposed an improved variant that uses a maximum filter on the frequency axis to suppress *vibrato* (i.e. a quasi-periodic variation in pitch [23]), and formulates ODF as the rectified l_1 difference between the pre-processed (i.e. filtered) spectra between the current frame and a previous frame that is μ frames apart. As the overlap of the two frames under comparison is smaller, the peaks can be sharper in the resulting ODF. This *SuperFlux* method effectively reduces the number of false positives originating from vibrato and performs well in MIREX 2013 and 2014 [15].

This letter extends the SF method by taking into account more temporal information while formulating the ODF. Specifically, while the SF method measures the "audio novelty" (i.e. value in the ODF) [1] of a given frame by comparing it with *one* of its previous frames, the proposed method measures the audio novelty by the *reconstruction error* while we use the linear combination of *a collection* of the previous frames to approximate the given frame. In other words, we assume that an onset event is the time instance when the current frame cannot be easily predicted by its previous frames. In this way, the computation of the ODF becomes a convex optimization problem, where additional constraints, such as the sparsity or non-negativity of the combination coefficients, can be added. We refer to this as a *linear reconstruction* (LR) method.

Ideas of using multiple frames in onset detection can be found in existing methods such as high-order linear predictive modeling [24]–[26] or bidirectional long short-term memory RNN [10]. The proposed method differs from the prior arts in two ways. First, we use all components as a whole spectral vector for linear reconstruction, while the prior arts perform linear prediction along specific bands of one-dimensional time samples independently and then accumulate the band-wise residual as

Manuscript received March 04, 2015; revised June 12, 2015; accepted July 28, 2015. Date of publication August 11, 2015; date of current version August 13, 2015. This work was supported by the Academia Sinica Career Development Program under Grant 102-CDA-M09. The associate editor coordinating the review of this manuscript was Dr. Zhu Liu.

The authors are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11564, Taiwan (e-mail: mister2dot4@gmail.com; lisu@citi.sinica.edu.tw; yang@citi.sinica.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2015.2466447

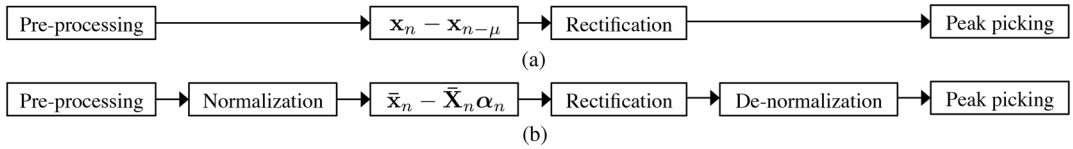

Fig. 1. The flowcharts of (a) the state-of-the-art SuperFlux method [15] and (b) the proposed linear reconstruction method for musical onset detection.

the total prediction error. Second, the LR method is optimization-based and therefore we propose and compare four possible implementations for musical onset detection, which has not been done elsewhere.

Through empirical evaluations, we show that LR can outperform SuperFlux when non-negativity constraint is imposed. Moreover, due to the longer temporal context involved, we find that the proposed LR method is in particular effective for *soft-onset instruments* (e.g. bowed string instruments) that exhibit rich vibrato and soft onsets. Due to the long attack phase with a slow rise in energy, such soft onsets can easily lead to false negatives for existing methods.

II. Method

The proposed method and SuperFlux only differ in the way the ODF is calculated. Other than that, the two methods use the same pre-processing and peak picking algorithms. We show their flowcharts in Fig. 1 and present the details below. In what follows, we use bold upper case and bold lower case to represent matrices and column vectors, respectively.

A. Pre-processing

Given a music signal sampled at 44,100 Hz, we first use the short-time Fourier transform to compute the magnitude spectra with a frame length of 2,048 points and a frame rate of 200 frames-per-second. Following [15], the magnitude Fourier spectra are then processed by 141 triangular filter banks ranging from 30 Hz to 17,000 Hz with the interval of 24 bands per octave. The resulting 141-dimensional feature vector per frame is then mapped to the logarithm scale by the mapping function $z \leftarrow \log(1+z)$ resulting in the auditory spectral feature vector $\mathbf{y}_n \in \mathbb{R}^K$[17], where n indexes time and $K = 141$ is the feature dimension. Finally, each feature vector is processed by a maximum filter with three bands of width along the frequency axis to suppress vibrato. We denote the resulting maximum-filtered feature vector as $\mathbf{x}_n \in \mathbb{R}^K$.

B. ODF Calculation: SuperFlux

SuperFlux takes the l_1-norm difference between two frames as the ODF, using the following formulation [13]:

$$\mathrm{ODF}_{\mathrm{SuperFlux}}(n) = \|h(\mathbf{y}_n - \mathbf{x}_{n-\mu})_+\|_1, \tag{1}$$

where $\mu > 0$ is an integer indicating the temporal offset, $\|\mathbf{z}\|_1 = \sum_{k=1}^{K} |z_k|$ is the l_1 norm, and $h(\mathbf{z}) = \frac{1}{2}(\mathbf{z} + |\mathbf{z}|)$ is the rectifier function. The rectifier function returns exactly the value of the difference if the magnitude of the frequency band increases from $n-\mu$ to n, and 0 if the magnitude decreases. The ODF is therefore the sum of the positive difference of every two feature vectors with μ frames apart along time.

C. ODF Calculation: The Proposed LR Method

Instead of merely comparing the spectral difference between two frames, we consider up to τ previous frames in calculating the ODF via the following linear reconstruction (LR) problem.

$$\begin{aligned}\{\mathbf{r}_n^*, \boldsymbol{\alpha}_n^*\} &= \arg\min_{\mathbf{r}_n, \boldsymbol{\alpha}_n} \|\mathbf{r}_n\|_2^2 + \lambda \cdot g(\boldsymbol{\alpha}_n), \\ \mathbf{r}_n &= \bar{\mathbf{x}}_n - \bar{\mathbf{X}}_n\boldsymbol{\alpha}_n,\end{aligned} \tag{2}$$

where $\boldsymbol{\alpha}_n \in \mathbb{R}^\tau$ denotes the combination coefficients of the τ previous frames $\bar{\mathbf{X}}_n \in \mathbb{R}^{K\times\tau}$ to reconstruct the current frame, $\|\mathbf{z}\|_2^2 = \sum_{k=1}^{K} z_k^2$ is the squared l_2 norm, and $g(\cdot)$ denotes the regularization term penalized by λ. To solve the problem (2), we require that all the input feature vectors to be l_2-normalized beforehand [27], [28]. In consequence, $\bar{\mathbf{x}}_n = \mathbf{x}_n/\|\mathbf{x}_n\|_2$, and $\bar{\mathbf{X}}_n = [\bar{\mathbf{x}}_{n-\mu}, \bar{\mathbf{x}}_{n-\mu-1}, \ldots \bar{\mathbf{x}}_{n-\mu-(\tau-1)}]$.

The parameter τ is referred to as the *reconstruction length*, and it should be a non-negative integer. Moreover, the residual $\|\mathbf{r}_n\|_2 = \|\bar{\mathbf{x}}_n - \bar{\mathbf{X}}_n\boldsymbol{\alpha}_n\|_2$ is viewed as the reconstruction error and is expected to be indicative of onset events.

We consider the following four implementations of (2)

- **Ordinary least square** (OLS): $\lambda = 0$ and no regularizers.
- **Non-negative least square** (NNLS): $\lambda \to \infty$ and $g(\boldsymbol{\alpha}_n) = \sum_{i=1}^{\tau} (\alpha_{ni})_-$, where α_{ni} denotes the i-th element of $\boldsymbol{\alpha}_n$ and the function $(z)_-$ returns 1 if $z < 0$ and 0 otherwise. In other words, $g(\boldsymbol{\alpha}_n) = 0$ if and only if all the elements in $\boldsymbol{\alpha}_n$ are non-negative. With this constraint, we ensure that the reconstruction is only additive.
- **Basis pursuit denoising** (BPDN): $\lambda > 0$ and $g(\boldsymbol{\alpha}_n) = \|\boldsymbol{\alpha}_n\|_1$. This formulation can usually lead to a sparse solution of $\boldsymbol{\alpha}_n$[29], meaning only a few non-zero elements. This sparsity constraint implies only a subset of previous frames is used to reconstruct a given frame.
- **BPDN with the non-negativity constraint** (BPDN+): the combination of the above two, considering both the non-negativity and sparsity constraints.

The first two methods, OLS and NNLS [30], are implemented with the 'numpy.linalg.lstsq' and 'scipy.optimize.nnls' functions of open-source Python libraries [31], [32]. The last two can be addressed by many optimization algorithms [33]–[35]. Our implementation uses the homotopy-based least angle regression and shrinkage (LARS)-Lasso algorithm [36] from the open-source toolbox SPAMS [37], for its demonstrated efficiency and effectiveness [38]. We set $\lambda = 0.001$ empirically.

Instead of directly using $\|\mathbf{r}_n\|_2$ as the ODF, we further incorporate the idea of rectification and calculate

$$\mathrm{ODF}_{LR}(n) = \|\mathbf{r}_n \odot (\mathbf{x}_n - \mathbf{x}_{n-\mu})_+\|_2 \cdot \|\mathbf{x}_n\|_2, \tag{3}$$

where $\odot$ is the element-wise product and $(x)_+ = \max(x, 0)$ is the element-wise rectification operator. That is to say, only the frequency bands with increased energy from $n-\mu$ to n in the

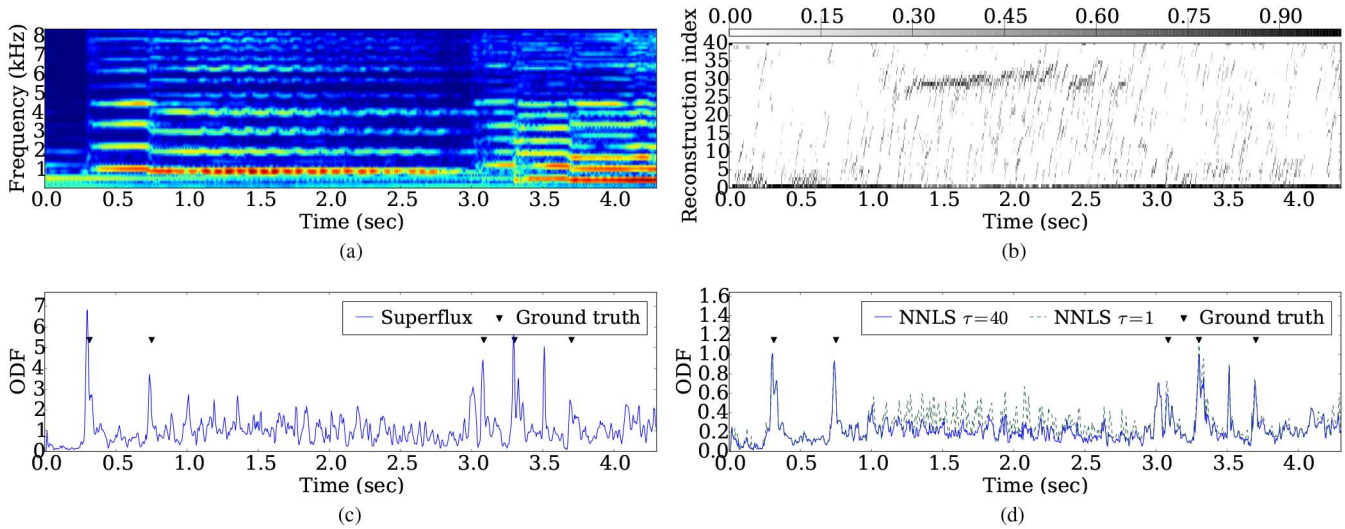

Fig. 2. (a) The pre-processed spectrogram of a bow-string cello sample comprising challenging cases including soft onset and vibrato. Vibrato can be found in 1.06–3.03 second. (b) The reconstruction coefficients $\boldsymbol{\alpha}_t$ of the NNLS method across time, where the black regions represent the activation of basis among the previous μ-th frame (bottom) to the previous $(\mu+\tau-1)$-th frame, with $\mu = 3$ and $\tau = 40$. (c) The ODF of SuperFlux, with the groundtruth onsets indicated by downward triangles. (d) The ODFs of NNLS using different reconstruction lengths: $\tau = 1$ (dark green dash line) and $\tau = 40$ (blue line).(a) Magnitude spectrogram of a music signal (b) Reconstruction coefficients obtained by NNLS (c) ODF of Superflux, a spectral flux method (d) ODF of NNLS, a linear reconstruction method.

original, un-normalized feature vectors are considered in calculating the reconstruction error. Moreover, the rectified reconstruction error is multiplied by the l_2 norm of the original feature vector $\mathbf{x}_n$. This *de-normalization* suppresses the frames of weak energy and empirically improves the performance of the proposed method, as will be shown in Section III.

D. Peak Picking

Following the settings of the SuperFlux-based submission to MIREX 2014 [15], we consider the following three heuristics in picking onset candidates from the ODF, for both SuperFlux and the proposed method. There is an onset at time t if,

- the ODF at time t has the maximal magnitude over the window from $t - 10$ ms to $t + 50$ ms,
- and, the ODF at time t has magnitude greater than or equal to a threshold δ plus the average magnitude of the ODF over the window from $t - 150$ ms to t,
- and, there is no other detected onsets in the last 30 ms.

These parameters can be fine-tuned, e.g. we can consider the window from $t-30$ ms to t in the first heuristic for online applications (i.e. having no access to future information) [15]. However, for simplicity, we only empirically tune the value of the threshold δ in the second heuristic, since the magnitude range of the ODFs can be different. We finally set δ to 1.5 and 0.3 for SuperFlux and the proposed method (including the four possible implementations), respectively, after optimizing the value for the overall dataset (see Section III). Moreover, we optimize the value of μ and set $\mu = 3$ for both methods.

E. Example

To gain insights, we show in Fig. 2 the ODFs of one cello solo, '14_ViolonceloTaksim_pt1,' selected from the experimental dataset [3]. We can see from Figs. 2(a) and (c) that the piece contains rich vibrato components that cannot be effectively suppressed by maximum filter and that would cause unfavorable prominent peaks (and accordingly false positives) in $\text{ODF}_{\text{SuperFlux}}$ (e.g. in $1.06-3.03$ seconds). Fig. 2(d) shows that, NNLS can effectively mitigate vibrato when the reconstruction length is sufficiently long (i.e. $\tau = 40$), without compromising the prominence of the peaks for the true onsets. The prominence of some onsets (e.g. the one at 0.8 second) is even enhanced. We see from Fig. 2(a) that the note from $0.86-3.03$ seconds has a soft attack time of about 0.20 seconds and vibrato rate of about 6 Hz. Even for such a case, with $\tau = 40$ (equivalent to 0.20 seconds) the proposed method has enough temporal context to model the temporal fluctuation and to disregard the vibrato.

Interestingly, the activation frames can be observed from the reconstruction coefficients $\boldsymbol{\alpha}_n$ displayed in Fig. 2(b). Not surprisingly, we see many non-zero elements in the bottom row, suggesting that the closest frame is often selected in the reconstruction of a given frame. However, we also see clusters of non-zero elements around the 30th previous frame, equivalently 150 ms earlier than the current frame. This is near to one period of vibrato (e.g. 6 Hz). While the maximum filter in SuperFlux operates along the frequency axis, our method can be viewed as a *temporal filter* that suppresses the unwanted temporal fluctuation and timbre variation in music.

Another interesting observation is that, although we do not enforce sparsity for NNLS, it appears from Fig. 2(b) that the activation pattern of NNLS is fairly sparse. This may due to the requirement of additive reconstruction set forth by the non-negativity constraint. We will show in Section III that NNLS is the most effective one among the four LR methods.

III. Experiment

We evaluate the performance of onset detection using the dataset compiled by Holzapfel *et al.* [3], which is composed of 2,169 onset events for 11 categories of monophonic instruments (ney, cello, violin, tanbur, piano, saxophone, kemençe, clarinet, trumpet, oud, and guitar) plus a category of polyphonic mixture. The dataset is a challenging one as it contains bowed-string

TABLE II
F-SCORE AND EFFICIENCY COMPARISON BETWEEN SUPERFLUX AND THE PROPOSED NNLS METHOD. BOLD FACE INDICATES BETTER RESULT

	ney	cel	vio	tan	pia	sax	kem	cla	mix	tru	oud	gui	Overall	Time (s)
SuperFlux	0.531	0.581	0.805	0.756	0.892	0.762	0.705	0.883	**0.839**	**0.881**	**0.923**	**0.824**	0.793	**2.93**
NNLS	**0.679**	**0.664**	**0.875**	**0.808**	**0.931**	**0.795**	**0.718**	**0.890**	0.838	0.871	0.908	0.797	**0.822**	5.88

TABLE I
RESULT OF ONSET DETECTION FOR LR METHODS, WITH DIFFERENT RECONSTRUCTION LENGTHS τ. BOLD FACE INDICATES BEST RESULT

	OLS	NNLS	BPDN	BPDN+	NNLS	
	F-score				Precision	Recall
$\tau = 1$	0.809	0.809	**0.809**	0.809	0.778	0.844
$\tau = 3$	**0.811**	0.818	0.808	0.815	0.798	0.844
$\tau = 5$	0.807	**0.822**	0.806	**0.819**	0.806	0.839
$\tau = 10$	0.801	0.820	0.799	0.816	0.814	0.826
$\tau = 30$	0.744	0.811	0.767	0.807	0.839	0.785
$\tau = 50$	0.683	0.802	0.734	0.798	0.842	0.766

TABLE III
PERFORMANCE COMPARISON OF SUPERFLUX AND NNLS, AND NNLS WITH ONE OF THE THREE KEY BUILDING BLOCKS LEFT OUT

Method	F-score	Precision	Recall
SuperFlux ($\delta = 1.5$)	0.793	0.801	0.786
NNLS ($\delta = 0.3$)	0.822	0.806	0.839
NNLS, w/o normalization ($\delta = 1.2$)	0.743	0.662	0.845
NNLS, w/o rectification ($\delta = 0.3$)	0.704	0.593	0.865
NNLS, w/o de-normalization ($\delta = 0.06$)	0.714	0.611	0.859

instruments and flutes (i.e. ney) that have lots of soft onsets, vibrato and timbre fluctuation. The performance is evaluated in terms of precision, recall and F-score, using the mir_eval toolbox [39]. Following MIREX, for every groundtruth onset, we count only one of the predicted onsets, if any, that fall within a tolerance window of ± 50 ms around the groundtruth onset as a correct detection [12].

Table I tabulates the F-scores of the four LR methods with varying reconstruction lengths. By comparing the rows, we see that better results are obtained with a moderate value of τ. Although not completely shown in the table, as τ increases, the precision increases and the recall decreases for all the methods, possibly because using more previous frames makes it easy to reconstruct both onset and non-onset events. By comparing the columns, we see NNLS and BPDN+ perform relatively better than the remaining two, suggesting the importance of considering non-negativity constraint rather than sparsity constraint for this method. As long as non-negativity is imposed, we can actually drop the sparsity constraint. The best results are obtained by using the NNLS method, with $\tau = 5$.

Table II compares the overall and per-instrument F-scores of the proposed NNLS method ($\tau = 5$) and our implementation of SuperFlux based on the open-source codes shared by Böck and Widmer [13]. Instrument classes are labeled according to their first three alphabets (e.g. cel = cello, gui = guitar). We can see from Table II that NNLS outperforms SuperFlux for many instruments, especially for the soft-onset ones (e.g. ney, cello and violin). The overall F-score of the two methods are 0.822 and 0.793, respectively, which exhibits a significant difference (p-value < 0.001) under the one-tailed t-test.

Table II also shows that, while the Holzapfel's dataset lasts for 12.98 minutes in total, it only takes 5.88 seconds for NNLS to perform onset detection for the whole dataset, with i7 quad cores running at 3.3 GHz on a Mac mini Server. This computational time is only two times of SuperFlux. Unlike complicated methods such as bidirectional long short-term memory RNN [10], both SuperFlux and NNLS are applicable to online real-time scenarios with proper peak picking methods that do not use future information.

Table III lists the precision and recall of SuperFlux and NNLS. It can be seen that NNLS performs better mainly due to higher recall. Table III also shows the result when we leave out one of the following building blocks of the LR method: normalization, rectification, and de-normalization. Without either normalization or rectification, we have slightly higher recall, but much lower precision. The precision is particularly low if we drop out the normalization step. De-normalization is also important; without it the F-score is lower than that of the complete system by 0.108, even after we optimize the threshold δ of peak picking to 0.06. Without normalization and de-normalization, the proposed method would almost reduce to SuperFlux when $\tau = 1$.

We have also examined the reconstruction coefficients $\boldsymbol{\alpha}_t$ of NNLS using $\tau = 40$ (see Fig. 2(b)) and found that:

- Cello and violin are the only two instruments with salient activation around the 30th previous frame, which roughly corresponds to the vibrato components in music.
- The other instruments have most activation within the 10 closest frames. In particular, guitar, piano and clarinet seldom uses distant frames in the linear reconstruction.

These observations partially explain why $\tau = 5$ works well (cf. Table I), and, more importantly, suggest that it is possible to devise an instrument-dependent mechanism to set the reconstruction length τ for better result. This can be done, for example, by recognizing the instruments first [40]–[44] before performing onset detection. This is left as a future work.

IV. CONCLUSION

In this letter, we have presented a comprehensive evaluation of a novel constrained linear reconstruction (LR) based method to unsupervised musical onset detection. The proposed method is idea-wise simple, computationally light, and is more effective than the state-of-the-art SuperFlux method for instruments with soft onsets. We show that the proposed method is a multi-frame extension of the conventional spectral flux method, and validate by experiments the importance of the non-negativity constraint and rectification in the proposed method. We see three interesting future extensions: instrument-dependent onset detection, incorporation of other features such as group delay and pitch [3], [14], [19], and refinement of the reconstruction matrix $\bar{\mathbf{X}}_n$ by for example online dictionary learning [37].

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 Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, 2005.

[2] S. Dixon, "Onset detection revisited," in *Proc. Int. Conf. Digital Audio Effects*, 2006, pp. 133–137, Citeseer.

[3] A. Holzapfel, Y. Stylianou, A. C. Gedik, and B. Bozkurt, "Three dimensions of pitched instrument onset detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1517–1527, 2010.

[4] N. Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *Proc. Int. AES Conf.: Semantic Audio*, 2005.

[5] O. Lartillot, Z. F. Yazc, and E. Mungan, "A more informative segmentation model, empirically compared with state of the art on traditional turkish music," in *Proc. Int. Workshop on Folk Music Analysis*, 2013, vol. 63.

[6] H. von Coler and A. Lerch, "CMMSD: A data set for note-level segmentation of monophonic music," in *Proc. Int. AES Conf.: Semantic Audio*, 2014.

[7] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Challenges and futures directions," *J. Intell. Inf. Syst.*, vol. 41, no. 3, pp. 407–434, 2013.

[8] B. McFee and D. P. Ellis, "Better beat tracking through robust onset aggregation," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 2154–2158.

[9] A. Robertson and M. Plumbley, "B-Keeper: A beat-tracker for live performance," in *Proc. ACM Int. Conf. New Interfaces for Musical Expression*, 2007, pp. 234–237.

[10] F. Eyben, S. Böck, B. Schuller, and A. Graves, "Universal onset detection with bidirectional long short-term memory neural networks," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2010, pp. 589–594.

[11] S. Böck, A. Arzt, F. Krebs, and M. Schedl, "Online real-time onset detection with recurrent neural networks," in *Proc. Int. Conf. Digital Audio Effects*, 2012, pp. 1–4.

[12] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2012, pp. 49–54.

[13] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Proc. Int. Conf. Digital Audio Effects*, 2013 [Online]. Available: https://github.com/CPJKU/SuperFlux

[14] S. Böck and G. Widmer, "Local group delay based vibrato and tremolo suppression for onset detection," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2013, pp. 361–366.

[15] S. Böck, F. Korzeniowski, and F. Krebs, "MIREX 2014 submissions," *Music Information Retrieval Evaluation eXchange*, 2014 [Online]. Available: http://www.music-ir.org/mirex/abstracts/2014/SB2.pdf

[16] J. Schlüter and S. Böck, "Improved musical onset detection with convolutional neural networks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 6979–6983.

[17] E. Marchi, G. Ferroni, F. Eyben, L. Gabrielli, S. Squartini, and B. Schuller, "Multi-resolution linear prediction based features for audio onset detection with bidirectional LSTM neural networks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 2164–2168.

[18] Z. Zhang, D. yan Huang, R. Zhao, and M. Dong, "Onset detection based on fusion of SIMPLS and superflux," *Music Information Retrieval Evaluation eXchange*, 2013 [Online]. Available: http://www.music-ir.org/mirex/abstracts/2013/ZHZD1.pdf

[19] L. Su and Y.-H. Yang, "Power-scaled spectral flux and peak-valley group-delay methods for robust musical onset detection," in *Proc. Sound and Music Computing Conf.*, 2014.

[20] M. Tian, G. Fazekas, D. A. A. Black, and M. Sandler, "Design and evaluation of onset detectors using different fusion policies," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2014, pp. 631–636.

[21] J. S. Downie, X. Hu, J. H. Lee, K. Choi, S. J. Cunningham, and Y. Hao, "Ten years of MIREX (music information retrieval evaluation exchange): Reflections, challenges and opportunities," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2014, pp. 657–662 [Online]. Available: http://www.music-ir.org/mirex/wiki/MIREX_HOME

[22] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *Proc. Int. Conf. Digital Audio Effects*, 2002, pp. 33–38.

[23] M. Mellody and G. H. Wakefield, "The time-frequency characteristics of violin vibrato: Modal distribution analysis and synthesisa," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 598–611, 2000.

[24] W.-C. Lee and C.-C. J. Kuo, "Musical onset detection based on adaptive linear prediction," in *Proc. IEEE Int. Conf. Multimedia and Expo.*, 2006, pp. 957–960.

[25] W.-C. Lee and C.-C. J. Kuo, "Improved linear prediction technique for musical onset detection," in *IIH-MSP'06. Int. Conf. Intelligent Information Hiding and Multimedia Signal Processing, 2006*, 2006, pp. 533–536, IEEE.

[26] J. Glover, V. Lazzarini, and J. Timoney, "Real-time detection of musical onsets with linear prediction and sinusoidal modeling," *EURASIP Adv. Signal Process.*, no. 1, pp. 1–13, 2011.

[27] D. W. Marquardt, "Comment: You should standardize the predictor variables in your regression models," *J. Amer. Statist. Assoc.*, vol. 75, no. 369, pp. 87–91, 1980.

[28] P. O. Hoyer, "Non-negative sparse coding," in *Proc. IEEE Workshop Neural Networks for Signal Processing*, 2002, pp. 557–565.

[29] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, pp. 797–829, 2006.

[30] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*. Philadelphia, PA, USA: SIAM, 1974, vol. 161.

[31] T. E. Oliphant, "Python for scientific computing," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 10–20, 2007.

[32] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, "The numpy array: A structure for efficient numerical computation," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, 2011.

[33] A. Y. Yang, S. S. Sastry, A. Ganesh, and Y. Ma, Fast l_1-minimization algorithms and an application in robust face recognition: A review Univ. Illinois at Urbana-Champaign, Urbana, IL, USA, Tech. Rep., 2010.

[34] T. Mingkui, W. T. Ivor, and W. Li, "Matching pursuit LASSO part I: Sparse recovery over big dictionary," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 727–741, 2015.

[35] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Found. Trends Mach. Learn.*, vol. 4, pp. 1–106, 2012.

[36] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani *et al.*, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[37] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. Int. Conf. Machine Learning*, 2009, pp. 689–696 [Online]. Available: http://spams-devel.gforge.inria.fr

[38] C.-C. M. Yeh, P.-K. Jao, and Y.-H. Yang, The AWtoolbox for characterizing audio information Academia Sinica, Beijing, China, Tech. Rep. TR-CITI-15-001, 2015 [Online]. Available: http://mac.citi.sinica.edu.tw/awtoolbox/

[39] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. Ellis, "mir_eval: A transparent implementation of common MIR metrics," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2014, pp. 367–372.

[40] F. Fuhrmann and P. Herrera, "Quantifying the relevance of locally extrated information for musical instrument recognition from entire pieces of music," in *Proc. Int. Soc. Music Information Retrieval*, 2011, pp. 239–244.

[41] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2012, pp. 559–564.

[42] L.-F. Yu, L. Su, and Y.-H. Yang, "Sparse cepstral codes and power scale for instrument identification," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 7460–7464.

[43] D. Giannoulis, E. Benetos, A. Klapuri, and M. D. Plumbley, "Improving instrument recognition in polyphonic music through system integration," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 5259–5263.

[44] Z. Duan, B. Pardo, and L. Daudet, "A novel cepstral representation for timbre modeling of sound sources in polyphonic mixtures," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 7545–7549.

This paper previously published in IEEE Signal Processing Letters