Guitar note onset detection based on a spectral sparsity measure

Mina Mounir, Peter Karsmakers, and Toon van Waterschoot
KU Leuven, Department of Electrical Engineering (ESAT)
(1) ESAT-ETC, AdvISe Lab, Kleinhoefstraat 4, 2440 Geel, Belgium;
(2) ESAT-STADIUS, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
Email: {mina.mounir,peter.karsmakers,toon.vanwaterschoot}@esat.kuleuven.be

***Abstract*—The detection of note onsets is gaining a growing interest in audio signal processing research due to its wide range of applications in music information retrieval. We propose a new note onset detection algorithm NINOS^2 exploiting the spectral sparsity difference between different parts of a musical note. When compared to the popular state-of-the-art *LogFiltSpecFlux* algorithm, the proposed algorithm shows up to 61% better performance for automatically annotated guitar melodies as well as chord progressions. We also propose an additional performance measure to assess the relative position of detected onsets w.r.t. each other.**

I. Introduction

A note onset is conceptually defined as the time instant at which a musical note is played. Looking at the musical signal nature where a note is decomposed into a transient followed by a steady-state portion [1], onsets are points chosen to be as close as possible to the start of transients (attacks) [2]. In this paper, the proposed methodology for onset detection is based mainly on the definition found in [3] where an onset is defined as the first detectable part of the note in an isolated recording.

There is a growing interest in defining and detecting onsets, not only because transients play an important role in timbre perception [4], hence in instrument identification, but also because onset detection is useful in a wide range of applications: automatic music transcription [5], sound analysis (tempo and beat tracking) [6] and synthesis (enhancement of attacks) [7], and adaptive audio effects (time stretching). Moreover, the detection of note onsets is of importance in the growing field of *music information retrieval* and its added value for music search engines and recommender systems.

In the note onset detection literature, different definitions and models for signal onsets have been used, different ways of (often manually) labeling onsets have been employed to generate the ground truth for performance evaluation, and several algorithms capable of automatically detecting onsets up to a certain precision have been proposed. A general scheme for onset detection algorithms was introduced by [2]. The main component of this scheme is the *reduction function* or *onset detection function (ODF)* defined as a highly subsampled version of the original music signal having distinguishable amplitude peaks at time instants where onsets appear [3].

Existing methods for onset detection can be classified into two main groups: probabilistic and non-probabilistic methods. Whereas in the first methods, a probabilistic model [8] is learned or a neural network [9] is trained and then ODFs are calculated, in the latter methods ODFs are calculated directly from the signal or its extracted features [4], [10]. Even though the probabilistic methods seem to outperform the non-probabilistic ones by a small factor [11], the former need to be trained on large data sets to achieve more generic results. It is important to note that when using learning methods the features are learned but do not necessarily have good correspondence with musical properties, as opposed to the features typically used with non-probabilistic methods [4].

Many non-probabilistic solutions can be found in the literature [1] and differ in the signal representation used in the algorithm: time domain amplitude or frequency domain amplitude, phase, or both [12]. These solutions also differ in the type of signal operations used to compute the ODF: energy magnitude, distribution, derivative ...etc [4]. By looking at the MIREX note onset detection results [11], the state-of-the-art non-probabilistic method is the *ComplexFlux* method [13] which adds the phase information to the *SuperFlux* [10] method. Both methods aim to solve a special case (vibrato and tremolo suppression) and are based on the*LogFiltSpecFlux* [12] method which applies some pre- and post-processing to the *Detection by spectral dissimilarity* or *SpectralFlux* method proposed earlier in [4]. These methods all share the same basic idea, i.e. to detect onsets by looking at the temporal evolution of the magnitude spectrogram. The SpectralFlux ODF denoted Δ is given by:

$$\Delta(i) \ = \sum_{k=1}^{k=\frac{N}{2}} H(X_{ik} - X_{i-1,k}) \ , \qquad (1)$$

where $H(x) = \frac{x+|x|}{2}$ is the half-wave rectifier function and X_{ik} is the magnitude spectrogram for a music signal frame with frame index i and frequency bin k. The *LogFiltSpecFlux* ODF is obtained similarly but using the magnitude spectrogram coefficients' logarithm instead which results in a slight

This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of KU Leuven Impulse Fund IMP/14/037, the FP7-PEOPLE Marie Curie Initial Training Network "DREAMS" funded by the European Commission under Grant Agreement no. 316969, IWT O&O Project nr. 150432, and IWT O&O Project nr. 150611. The scientific responsibility is assumed by its authors.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

onset detection improvement. From the above definition, these ODFs would perform poorly with consecutive notes sharing a considerable amount of harmonics or with repeated notes with insufficient increase in magnitude. Such note sequences would be highly present in melodies formed by chords sequences, where chords would have higher probability of shared harmonics.

This paper presents a new non-probabilistic method for onset detection termed INOS^2 (*Identifying Note Onsets based on Spectral Sparsity*), as well as its normalized version termed NINOS^2. The proposed ODF is a (normalized) sparsity measure of the magnitude spectrogram of each input signal frame. Even though sparsity is an important distinguishing feature between the magnitude spectrum of a note's transient and steady-state parts, it has not been explicitly used in existing approaches to note onset detection. Moreover, since the proposed ODF does not rely on spectral differences between successive signal frames, the INOS^2 method overcomes the previously discussed problems related to certain note and chord sequences.

Another challenging problem in note onset detection algorithm design, is to unambiguously define and calculate the onset ground truth values, as it is difficult to tell what a "correct definition" of onsets is. Most of the proposed methods are evaluated with datasets that are manually annotated by looking at signal waveforms and/or spectrograms and listening to signal recordings [3]. In this paper we introduce a different evaluation approach, in which synthetic music signals based on real, automatically annotated musical note recordings are generated, as will be explained in Section III-A.

Having introduced the problem, the related work and challenges, Section II will emphasize the concept and explain the details of the proposed onset detection algorithm. The experimental evaluation is shown in Section III comparing the NINOS^2 and *LogFiltSpecFlux* methods. Finally, Section IV presents the conclusion and hints for future work.

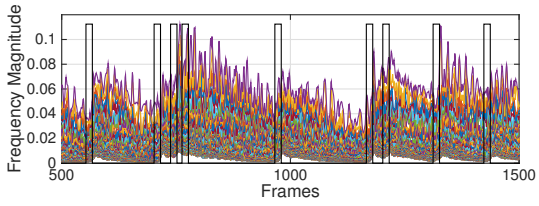
II. Proposed Note Onset Detection Method

State-of-the-art onset detection methods follow a certain scheme where the input signal undergoes four operations: pre-processing, ODF calculation, post-processing and peak-picking [2]. In the pre-processing step, the signal is filtered in order to emphasize some aspects or remove irrelevant noise making the detection easier. Then the signal is processed by a reduction algorithm in order to calculate the ODF. The resulting ODF may or may not undergo another filtering (post-processing) step before applying peak-picking to determine the onsets position in time. In this paper the pre-processing step is skipped and the focus will be on the ODF and the remaining operations.

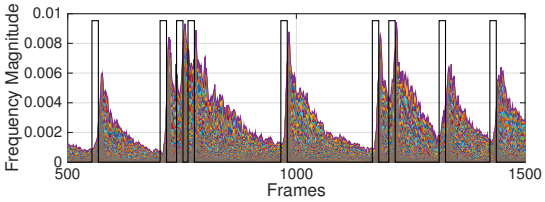
A. Proposed Onset Detection Function (ODF)

The proposed ODF is based on the fact that any musical note can be expressed as a sum of sinusoids. While the steady-state part of a note is well approximated by a small number of sinusoids, the transient part, being a short interval of time where the statistical and energy properties of the signal change rapidly [1], requires a much higher number of sinusoids to be accurately represented. Consequently, in the magnitude spectrogram of a musical note, the transient (attack) part of the note is spectrally less sparse than the following steady-state (tonal) part.

To maximize the differentiation between onsets and non-onsets in terms of spectral sparsity, we first select a subset of magnitude spectrogram coefficients. For guitars the note's fundamental frequency and harmonics have high energy during attacks and then decrease slowly, while the other frequency components are present almost only during attacks and generally have lower energy than the harmonic components. Fig. 1 makes it obvious that low-energy coefficients (1b) are more representative for onsets than high-energy ones (1a). Hence by ordering the magnitude spectrogram coefficients by their energy along the frequency dimension and removing the high-energy ones, thus neglecting fundamentals and harmonics, before applying the sparsity measure will enhance the discriminative power of the proposed ODF.



(a) High-energy coefficients.



(b) Low-energy coefficients.

Fig. 1. High- and low-energy coefficients behavior vs onsets.

We will now explain the details of how to calculated the INOS^2 ODF. First, the input signal is divided into overlapping windowed frames $x_1, x_2, \ldots, x_L$. For each frame x_i the magnitude spectrogram $X_{ik} \in \mathbb{R}^N$ is calculated using the *discrete Fourier transform (DFT)*,

$$X_{ik} = |\mathcal{F}(x_i)|\,, \quad i = 1, \ldots, L, \quad k = 1, \ldots, N. \quad (2)$$

Then the magnitude spectrum coefficients in each frame X_{ik} are sorted in ascending order and only the first J out of N coefficients are used afterwards, with

$$J = \lfloor\, \gamma N \,\rfloor, \quad J \in \mathbb{Z} \text{ and } 0 < \gamma < 1$$

where $\lfloor\cdot\rfloor$ denotes the floor function.

Finally, the INOS^2 ODF is calculated by measuring the spectral sparsity of the subset magnitude spectrogram X_{ij}. The ODF is chosen to be an inverse-sparsity measure as to have peaks for non-sparse frames, thus highlighting possible

onset locations in time. The INOS^2 ODF is an inverse-sparsity measure Υ defined per frame as:

$$\Upsilon(i) \;=\; \frac{\|X_i\|_2^2}{\|X_i\|_4} \;=\; \frac{\sum_{j=1}^{J} X_{ij}^2}{\left(\sum_{j=1}^{J} X_{ij}^4\right)^{\frac{1}{4}}} \;. \tag{3}$$

As stated in [14], to check whether or not a function could be used as a sparsity measure, it should satisfy two conditions which form together the sparsity definition. Firstly, the most sparse signal is the one with all its energy concentrated in one single coefficient. Secondly, the least sparse would be a signal having the energy distributed equally over all its coefficients. For example,

$$\begin{aligned} S_{max} &= S([0,0,0,0,1]) > \dots \\ &> S([0,0,1,1,1]) > \dots \\ &> S([1,1,1,1,1]) = S_{min} \end{aligned}$$

where S is a sparsity measure defined for vectors of equal length. By applying this conceptual definition, it can be easily shown that the INOS^2 ODF in (3) is an inverse-sparsity measure.

More specifically, Υ is a joint sparsity and energy measure. This becomes clear when rewriting (3) as

$$\Upsilon(i) \;=\; \|X_i\|_2 \cdot \frac{\|X_i\|_2}{\|X_i\|_4}. \tag{4}$$

The first term $\|X_i\|_2$ is the l_2-norm which represents the energy of the signal frame. This is a relevant feature for onset detection as usually onsets are accompanied with an energy rise. It has been used in the *envelope follower* method for onset detection [4]. On the other hand, the second term reflects sparsity. It is the ratio between the frame's l_2-norm and l_4-norm which increases as sparsity decreases. This is explained by applying the unit-ball concept [1] as shown in Fig. 2.

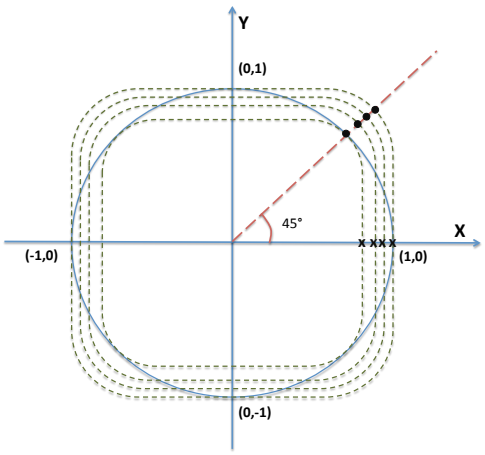


Fig. 2. Relation between l_2-norm and l_4-norm.

The figure shows a circle which represents the l_2-norm unit ball in $2D$. This is the set of points representing vectors having an l_2-norm equal to 1. First notice that being near the x or y axis, points represent sparse vectors and inversely while moving away from the axes. By applying the previously mentioned sparsity definition, the most sparse vectors are the points lying on the axes, i.e. $[0,1],[1,0],\dots$ etc., while the points lying on the $45°$ lines are the least sparse, e.g. $[\sqrt{0.5},\sqrt{0.5}]$. Focusing on the first quadrant, we can calculate the l_4-norm for each of the points on the l_2-norm unit ball. This could be graphically understood by looking at the l_4-norm unit ball –which is approximately a square– and its scaled versions each containing one of the points mentioned earlier. We observe that the scaled l_4-norm balls are getting smaller while moving from $[0,1]$ towards $[\sqrt{0.5},\sqrt{0.5}]$ and then bigger again while continuing to $[1,0]$. This means the l_2-norm is being larger than the l_4-norm when the vectors become less sparse and hence the ratio becomes larger.

Finally, by normalizing the sparsity measure to the number of coefficients, it can be shown that its inverse satisfies all of the desired sparsity measure criteria proposed in [14]. Using the empirical mean defined per frame X_i by

$$E_J\{X_i\} \;=\; \frac{1}{J}\sum_{j=1}^{J} X_{ij} \;,$$

the normalized ratio of norms is defined as

$$\frac{\sqrt[2]{E_J\{X_i^2\}}}{\sqrt[4]{E_J\{X_i^4\}}} \;=\; \frac{1}{\sqrt[4]{J}} \times \frac{\|X_i\|_2}{\|X_i\|_4} \;.$$

Hence the NINOS^2 ODF, i.e. the normalized version of the INOS^2 ODF, is defined as the normalized inverse-sparsity measure $\aleph$,

$$\aleph(i) \;=\; \frac{\|X_i\|_2^2}{\sqrt[4]{J}\|X_i\|_4} \;. \tag{5}$$

Even though the normalization does not much affect the onset detection results presented in this paper, it is necessary and useful in future work when processing frames having different lengths to obtain detections with different precisions.

B. Peak-Picking

For simplicity and fair comparison, the peak-picking used with the state-of-the-art algorithm [13] is applied in which a frame x_i is an onset candidate if all the following conditions are satisfied:

1) $\aleph(i) = \max_l\ \aleph(i+l), \quad \text{with} \quad l = -\alpha,\dots,+\beta,$
2) $\aleph(i) \geq \frac{1}{a+b+1}\sum_{l=-a}^{+b} \aleph(i+l) + \delta,$
3) $i - p > \Theta,$

where α, β, a, b and Θ are the peak-picking parameters: *before maximum*, *after maximum*, *before average*, *after average* and *combination width* counted in frame units, and p is the previous onset's index. An onset should be the highest ODF amplitude peak in its vicinity and is an amplitude offset δ above its neighborhood average. Finally an onset should be Θ frames apart from its predecessor in frame p.

While the peak-picking parameter values are kept the same as in [13], Θ is set equal to the detection window length

which is the maximum amount of frames in which a single ground-truth onset could occur. This value depends on the frame overlap and is calculated using the following relations:

$$\begin{aligned} h &= \lfloor (1-q)N \rceil, \\ r &= f_s/h, \\ \Theta &= \lceil rN/f_s \rceil. \end{aligned} \tag{6}$$

where $\lfloor \cdot \rceil$ and $\lceil \cdot \rceil$ denote the nearest integer and ceiling functions, h is the hop size in samples, q is the frame overlap factor from 0 to 1, N is the frame size in samples, r is the frame rate, f_s is the sampling frequency and Θ again is the amount of frames to be skipped after one onset detection before aiming to detect a new onset. This value should preferably be increased in case of instruments with very long attacks.

C. Onset Detection Algorithm Parameters

For a complete understanding of the (N)INOS^2 algorithm and of the performance measures explained later, we discuss some important algorithm parameters:

- *Processing frame size* (N): It should be larger than a single period of the signal [4] and small enough to capture transients.
- *Detection resolution*: It depends on the frame rate r which is inversely proportional to the hop size h.
- *Processing mode*: The detection algorithm could be run in either *offline* or *online* mode. In the latter, peak-picking parameters β and b are set to zero.
- *Ground-truth increase-factor*: It is used to increase the detection window to handle the lack of precision inherent in the ground-truth generation. Onsets may happen slightly before or after the ground-truth onset.

III. Experimental Evaluation

A. Data Set

As mentioned before, most of the annotated datasets for note onset detection are manually annotated [3], [10], [12]. Manual onsets annotation depends on many factors, e.g. the human visual and auditory accuracy or the musical note context (what comes before and after a note). In this paper we use an objective method for selecting onsets ground-truth, which is preferable to obtain a fair and accurate performance evaluation. To this end, we work with isolated notes or chords from the *McGill University Master Samples* library [15] and mix these to form a melody or chord progression. A software tool has been developed to load the different instrument notes, apply some amplitude effects (fade, loudness, etc.), and automatically annotate onsets and offsets (which can be easily and accurately done on isolated notes) depending on a short-term energy measure. Finally the annotated notes are mixed using some specifications (harmonic sequence, distances between onsets, etc.). In this way, artificial melodies and chord progressions are generated together with their automatically calculated ground-truth. This tool and the methodology are explained in more detail in [16].

We have tested the proposed algorithm with all guitar libraries found in [15], in order to cover acoustic and electric guitars, single notes and chords, as well as different playing styles. Only two guitar folders are excluded where the first "Electric Guitar Fifths" is facing poor detection performance with all tested algorithms and "Guitar Tapping" which contains percussive rather than pitched notes - body and not strings tapping. For each library, three test melodies are generated each having 50 notes: one test for algorithms tuning and the other two for algorithms evaluation. The temporal distances between notes are randomly chosen between 100 ms and 1 s. The input signals ($f_s = 44.1$ kHz) are divided into frames of $N = 2048$ samples (46 ms) with an overlap of $q = 90$ % leading to a detection resolution of 4.6 ms which is comparable to temporal hearing resolution ($\approx$ 10 ms). Then a 46 ms *Hanning* window is applied to each frame. The ground-truth increase-factor is set to 40 % which results in a 18.5 ms larger detection window in order to achieve the best average performance over all tuning experiments. In the tuning phase, the coefficients percentage γ is chosen to be 94 %.

B. Performance Measures

An important issue when assessing the detection algorithms is how true positives and negatives are counted. Here we adopted the same concept as the state-of-the-art algorithm where two onsets detected within one detection window are counted as one true and one false positive, and an onset counts only for one detection window [10].

The most common way to compare onset detection algorithms is by evaluating the corresponding F1-scores. We will compare results obtained with thresholds δ maximizing the F1-score for each algorithm. Since a true positive could occur anywhere within the detection window, a new measure is developed to determine how large the detection window should be in order to achieve the selected F1-score. This measure is defined per test melody by:

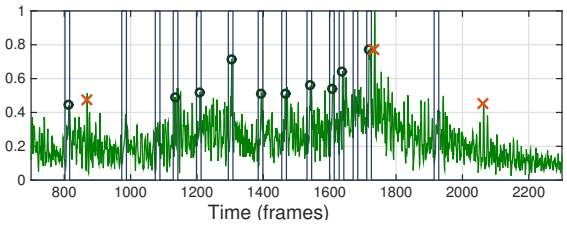
- *Detections mean* μ_d: The average onset relative position to the detection window start.
- *Detections standard deviation* σ_d: The corresponding standard deviation.

When σ_d is small, the algorithm will be detecting onsets in the same relative position to the start of the detection window. This is an important measure that reflects how well the algorithm detects the different onsets' relative position to each other.

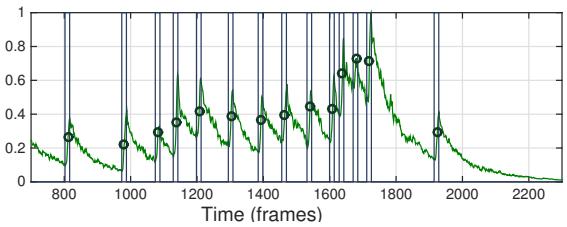
C. Results and Discussion

First, we discuss detection results and NINOS^2 performance compared to *LogFiltSpecFlux* (LSF) when applied on the guitar dataset. We first compare the respective ODFs. Figure 3 shows the two ODFs calculated for a short electric guitar melody having 14 different chords. Onsets ground-truth are marked with vertical lines and every two successive lines represent a single detection window. The peak-picking results are marked with circles for true positives and crosses for false positives, while false negatives are easily noted by unmarked detection windows.

While both ODFs present higher amplitudes at onsets, it is clear that the proposed ODF is smoother and hence



(a) LogFiltSpecFlux ODF



(b) NINOS^2 ODF

Fig. 3. Normalized ODFs for 14-chord guitar example.

facilitates the online peak-picking. This results in onsets being detected earlier and more precisely with the NINOS^2 method as compared to the *LogFiltSpecFlux* method. In the latter, because of the lack of ODF smoothness, the peak-picking typically detects onsets later and produces more false negatives due to the ripples in the ODF masking the onset peaks.

Next, detection performance is compared in Table I showing the average best-F1 score for each algorithm/library pair. For most of the analyzed melodies in the described dataset, the NINOS^2 method outperforms the *LogFiltSpecFlux* method except for some acoustic guitar melodies where both methods perform equally. Table I also compares the detection standard deviation σ_d showing again a better performance for the NINOS^2 method with the electric guitar and similar performance for the acoustic guitar.

TABLE I
PERFORMANCE MEASURES COMPARISON

	best-F1		σ_d	
	NINOS^2	LSF	NINOS^2	LSF
Acoustic Guitar				
Normal	1.0000	1.0000	2.2525	2.2927
Harmonics	0.9103	0.9031	2.5291	2.5145
Pizzicato	1.0000	1.0000	2.5248	2.3194
Sul Ponticello	1.0000	1.0000	1.5072	1.3020
Sul Tasto	0.9533	0.9277	2.5919	2.6788
Electric Guitar				
Normal	0.9764	0.9473	2.7229	2.8780
Dominant Ninth	1.0000	0.9499	1.3392	2.7099
Dominant Seventh	0.9641	0.8525	2.5659	3.3003
Elevenths	1.0000	0.9439	1.6265	2.6487
Flat 7 Sharp 9	0.9967	0.9763	1.4412	2.3409
Major Seventh	0.9900	0.7219	2.0706	4.6044
Major Seventh Stopped	0.8836	0.5922	2.9474	4.3673
Major Sixth	0.9439	0.5855	2.7527	4.5057
Major Seventh	0.9966	0.9471	1.4503	2.6040
Ninth	0.9933	0.8994	1.7907	3.1060
Harmonics	1.0000	0.9899	1.5658	1.5976
Stereo Chorus	0.9836	0.9765	1.6959	1.8370

IV. CONCLUSIONS AND FUTURE WORK

This paper introduced a new promising method for note onset detection based on spectral sparsity, showing up to 61 % better performance with guitar melodies and chord progressions. A new measure σ_d for the evaluation of note onset detection was proposed, emphasizing the importance of onsets relative distance. Moreover, our evaluation has been based on human-independent ground-truth annotation, making the comparison between different algorithms easier and more objective.

We are currently investigating a larger number of spectral sparsity measures that will be applied on different instruments with different playing styles. Because of the fact that the NINOS^2 ODF is constructed such as to fulfill general sparsity measure properties, it can be used in future work for analyzing melodies using a varying frame size.

REFERENCES

[1] M. Mounir, "Note onset detection using sparse over-complete representation of musical signals," Master's thesis, Advanced Learning and Research Institute - USI, 2013. [Online]. Available: ftp://ftp.esat.kuleuven.be/stadius/mshehata/mscthesis/mshehatamsc.pdf

[2] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 1035–1047, Sep. 2005.

[3] P. Leveau and L. Daudet, "Methodology and tools for the evaluation of automatic onset detection algorithms in music," in *Proc. Int. Symp. Music Information Retrieval*, 2004, pp. 72–75.

[4] P. Masri, "Computer modelling of sound for transformation and synthesis of musical signals," Ph.D. dissertation, University of Bristol, 1996.

[5] E. Benetos and S. Dixon, "Polyphonic music transcription using note onset and offset detection," in *2011 IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP)*, 2011, pp. 37–40.

[6] M. McKinney, D. Moelants, M. Davies, and A. Klapuri, "Evaluation of audio beat tracking and music tempo extraction algorithms," *J. New Music Research*, vol. 36, pp. 1–16, May 2007.

[7] P. Masri and A. Bateman, "Improved modelling of attack transients in music analysis-resynthesis," in *Proc. Int. Comput. Music Conf.*, 1996, pp. 100–103.

[8] N. Degara, M. Davies, A. Pena, and M. Plumbley, "Onset event decoding exploiting the rhythmic structure of polyphonic music," *IEEE J. Selected Topics Signal Process.*, vol. 5, pp. 1228–1239, Oct. 2011.

[9] J. Schluter and S. Böck, "Improved musical onset detection with convolutional neural networks," in *2014 IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP)*, 2014, pp. 6979–6983.

[10] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Proc. 16th Int. Conf. Digital Audio Effects (DAFx-13)*, 2013.

[11] "Mirex 2015 onset detection results," http://nema.lis.illinois.edu/nema_out/mirex2015/results/aod/, 2015, accessed 2015-12-08.

[12] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods," in *Proc. 13th Int. Soc. Music Information Retrieval Conf. (ISMIR)*, 2012, pp. 49–54.

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

[14] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Trans. Inf. Theory*, vol. 55, pp. 4723–4741, Oct. 2009.

[15] F. Opolko and J. Wapnick, *McGill University Master Samples*. Montreal, QC, Canada: McGill University, Oct 2006, DVD edition.

[16] M. Mounir and T. van Waterschoot, "New methodology for notes onset definition and detection based on spectral sparsity measures," KU Leuven, Belgium, Tech. Rep. ESAT-STADIUS TR 16-11, Feb. 2016.