Three Dimensions of Pitched Instrument Onset Detection

André Holzapfel, Yannis Stylianou, *Member, IEEE*, Ali C. Gedik, and Barış Bozkurt

***Abstract*—In this paper, we suggest a novel group delay based method for the onset detection of pitched instruments. It is proposed to approach the problem of onset detection by examining three dimensions separately: phase (i.e., group delay), magnitude and pitch. The evaluation of the suggested onset detectors for phase, pitch and magnitude is performed using a new publicly available and fully onset annotated database of monophonic recordings which is balanced in terms of included instruments and onset samples per instrument, while it contains different performance styles. Results show that the accuracy of onset detection depends on the type of instruments as well as on the style of performance. Combining the information contained in the three dimensions by means of a fusion at decision level leads to an improvement of onset detection by about 8% in terms of F-measure, compared to the best single dimension.**

***Index Terms*—Automatic music transcription, group delay, music information retrieval, onset detection.**

I. Introduction

IMAGINE yourself being in the audience of a folk music concert. A violin player starts playing a popular dancing tune, people start moving their feet to the rhythm. Then a guitar starts playing and people begin to clap their hands. The necessary condition that enables the human being to behave in this way is his ability to perceive the starting points of the musical notes. Onset detection, the detection of the starting instant of an event in a signal, is an extensively studied topic in various domains of signal processing. Musical onset detection, the detection of the starting point of a musical note transient [1], is one of the sub-domains with a very large literature. Various methods have been proposed, evaluations are taking place [2], and tutorials are available [1], [3].

The main challenge in a musical onset detection problem is to build a robust algorithm that can detect onsets of various types of signals, i.e., the notes of the tune played by the violin player as well as the accompaniment played by the guitar. Considering also the variations in musical styles (classical, pop, jazz to folk music, etc.,) and performance styles (playing with a pick, finger picking, ornamentation styles in folk musics, etc.) the variability is so large that it is problematic and time-consuming to collect a representative data set and to evaluate various methods comparatively.

The algorithmic steps of an onset detection system are as follows.

1) Preprocessing of the audio signal (the raw time-series data).
2) Computation of an onset strength signal (OSS), which is mainly a parameter time-series at a sampling frequency lower than that of the audio signal. The term OSS has been used in [4], while OSS are referred to as novelty function in [5] and as detection functions in [1].
3) Detection of transients in the OSS usually using a peak-picking algorithm [1].

The preprocessing is an optional step as in many other signal processing applications. The most common form of preprocessing used for musical onset detection is sub-band or multi-band decomposition. Most of the studies using multi-band processing, approach the problem in a similar fashion: dividing the signal into sub-bands, estimating OSS for each sub-band, combining either at the OSS level or the onset decision level to achieve a final decision for the onsets [6]–[10]. It is reported that the robustness is improved by using such methodology [7].

The core of the design of the onset detection system is the OSS estimation part for which a large variety of methods exist. Comprehensive reviews of these methods are available in [1], [3] and [11].

One type of approach for OSS computation is the use of temporal feature variations such as the time-domain amplitude envelope of the signal [12], short-time energy [6] or its variants [13]. These relatively old approaches are successful for processing clean recordings of instruments with percussive character; however, they have problems in processing, for example, bowed instrument sounds where musical note change does not always imply a sudden change in the energy or amplitude. Another common approach for OSS computation is the use of spectral features since spectrograms of recordings very often reveal clear visual clues of the onset locations. Due to the difficulties involved in phase processing, amplitude processing is much more common in spectral methods. Spectral flux, the amplitude spectra difference computed for consecutive frames using various distance functions (L-1 norm, L-2 norm, Kullback–Leibler distance, etc.), is used in many studies due to the simplicity of

Manuscript received January 09, 2009; revised October 11, 2009. First published November 17, 2009; current version published July 14, 2010. The work of A. C. Gedik and B. Bozkurt was supported by the Scientific and Technological Research Council of Turkey, TÜBITAK, under Project 107E024. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Patrick A. Naylor.

A. Holzapfel and Y. Stylianou are with the Institute of Computer Science, FORTH, GR-700 13 Heraklion, Crete, Greece, and the , Computer Science Department, Multimedia Informatics Laboratory, University of Crete, GR-714 09 Heraklion, Crete, Greece (e-mail: hannover@csd.uoc.gr; yannis@csd.uoc.gr).

A. C. Gedik and B. Bozkurt are with Electrical and Electronics Engineering Department, Izmir Institute of Technology, Urla, TR-35430 Izmir, Turkey (e-mail: a.cenkgedik@musicstudies.org; barisbozkurt@iyte.edu.tr).

Digital Object Identifier 10.1109/TASL.2009.2036298

1558-7916/$26.00 © 2010 IEEE

computation and robustness in detection of onsets of pitched-percussive sounds, for example in [9], [14], [11], [3].

Although less common, phase processing is also used for OSS computation. In [15] a phase-based OSS is presented for the first time. Their approach is based on the segmentation of a signal into transient and steady-state (TSS) frame-by-frame detecting fast instantaneous frequency changes. In [16], the previous phase based approach is improved by using a mean absolute phase deviation function or alternatively a difference function on the complex Fourier coefficients from consecutive short-time frames. Phase-based OSS are used in combination with energy-based detection functions in a number of other studies [9], [11], [3].

However, computing reliable phase deviation information (or similarly a complex Fourier coefficients deviation) from consecutive frames is problematic. The main problem is the phase unwrapping operation or window synchronization [17]. It has been shown in various studies ([17]–[19]) that a large number of very high jumps in the phase slope (i.e., the—negative—group delay function) of audio signals exist due to zeros of the z-transform closely located around the unit circle. Steiglitz and Dickinson [20] have shown that the roots of the z-transform of a short-time signal tend to be evenly distributed in angle and tightly clustered near the unit circle as the degree of the polynomial (length of the time domain signal) increases. Hence, zeros of the z-transform for 20–30 ms short-time audio signals are clustered around the unit circle resulting in many spikes in the phase slope [17]. This leads to the conclusion that reliable phase processing is very difficult to achieve unless certain synchronization rules (such as pitch synchronization in speech processing) are applied. The alternative to the direct usage of phase information is the processing of either some modified version of the group delay [21] or the average of group delay which can be used for detection of events like glottal closure instants (GCIs) [22], [23]. The average group delay has been applied for other types of transient detections as well, for example in the detection of clicks from marine mammals in [24]. Because detecting onsets in music is a transient detection problem as well, phase information can be used in a similar way as in click or GCI detection.

As the first contribution of this paper a new onset detection method is proposed. This method is based on processing the average of the group delay function which will be referred to as *phase slope function*. The derivative of phase along *time* is referred to as instantaneous frequency, and has been used in [15], among others, for onset detection. In this paper, the usage of group delay will be proposed, which is the derivative of phase along *frequency*. It is interesting to note that the observations made on phase plane plots in [25] showed that onsets appear more clearly when computing the derivative of the phase over frequency than over time. However, these observations were not developed further into any onset detection system. Recently, an onset detection method based on group delay was shown in [26] to improve the performance of beat tracking in music with little percussive content.

Another type of approach which specifically targets improvement of onset detection for non-percussive sounds is the usage of the fundamental frequency or pitch of the signal [27]. In [11], it was shown that previously presented approaches based on spectral features perform worse for pitched non-percussive than for pitched percussive sounds.

Onset detection is therefore performed using spectral amplitude, phase, and pitch information. These three features or cues define therefore a three-dimensional space.[1] We suggest that a human makes use of all these dimensions of the space for onset detection and the importance of each dimension (weight) depends on the type of musical signal. Thus, the second major contribution of this paper, beside the usage of group delay for onset detection, will be an appropriate combination of the information contained in these three dimensions. So far, only amplitude and phase information have been combined in various studies (for example [28]), where the phase information considers the instantaneous frequency changes and not the group delay as proposed in this paper. In [29], depending on the type of signal, either an energy based or a pitch-based detector is applied. In [30], statistical models are built for different features (MFCC, LPC coefficients and others), and the decisions derived from the different models are combined to a single decision function. To the best of our knowledge, it has not been tried yet to combine the three dimensions of pitch, spectral amplitude and phase in order to get an improved onset detector. In this paper, a combination of the decisions derived from the three individual dimensions (decision fusion) is proposed. This simple combination works without training complex statistical models for the feature distributions like in [30], and can easily be improved or extended by either changing one of the OSS or by adding a new one.

In order to determine the performance of different OSS, it is necessary to study the three feature dimensions and their fusion on a large enough dataset that is publicly available. The lack of common databases of pitched instruments is an important obstacle for further improvement. Thus, the third major contribution of this study is the compilation of such a publicly available database, and studying the above mentioned three dimensions on this dataset. Despite the fact that signal characteristics (hence the onset detection performance) vary largely for different types of instruments, very few studies include performance styles or instruments of traditional forms of music in their databases (for example in [31] and [32]). In our database, we intentionally include traditional Turkish music instruments (*ud*, *tanbur*, *ney*, and *kemençe*) to also study variations between western and non-western music. As a result, a data set containing a diverse set of pitched instruments is available for the evaluation of onset detection systems. The data set will be provided to interested researchers on request to the first author.

The content of this paper is organized as follows. Section II describes the data used for evaluation of the onset detection. Section III describes the computation of the three OSS for magnitude, phase slope, and fundamental frequency, along with the applied fusion method. Section IV gives details of the measures that are applied to evaluate the accuracy of the onset detection, that is obtained by applying the peak picking detailed in Section V to the proposed OSS. Section VI presents the experimental results, and Section VII concludes the paper.

[1]we use the term "dimension" freely without its formal definition

TABLE I
MAIN DATASET DETAILS (1)

Main Set (MS)		
Instrument	Number of Onsets	Number of files
cello	150	5
clarinet	149	5
guitar	174	5
kemençe	186	5
ney	147	7
ud	211	5
piano	195	5
saxophone	148	5
tanbur	156	5
trumpet	140	5
violin	173	5
Sum	1829	57

TABLE II
DEVELOPMENT DATASET DETAILS (2)

Development Set (DS)		
Instrument	Number of Onsets	Number of files
guitar	147	7
ud	207	5
piano	117	6
violin	203	3
Sum	674	21

II. DATA SET

In order to evaluate the performance of a musical instrument onset detector, an annotated dataset is necessary. Although in recent years many publications have treated the problem of onset detection, experiments are usually performed on small datasets with uneven class distributions [1], or on datasets containing samples with several instruments playing at the same time. No sufficiently large onset annotated dataset of different pitched musical instruments is publicly available. Such a dataset would make fundamental research on the accuracy of onset detection techniques feasible. For this paper, a dataset as described in Table I has been compiled, which will be referred to as the main data set (MS) in the following. Non-pitched percussive instruments, such as drums and percussions, have not been included in this data set as their onsets can be considered easy to detect, for example by using criteria derived from their energy envelope [1]. Since for onset detection the characteristic of the excitation is a crucial point, the instruments have been grouped into the following classes according to this aspect: pitched-percussive instruments (guitar, *ud*, piano and *tanbur*), wind instruments (clarinet, *ney*, saxophone and trumpet) and bowed string instruments (cello, *kemençe* and violin). All samples are monophonic. We also tried each of the above classes to be represented by a similar number of samples and instruments. Furthermore, besides the choice of instruments commonly used in western music, also instruments of Turkish music are included (*kemençe*, *ney*, *ud*, and *tanbur*). This enables us to compare the influence of the musical style on the accuracy of onset detection systems. The Turkish music examples were chosen in order to select samples that are representative for the style of performance but that do not contain many notes at which hand annotation would have been too error-prone. This restriction has been found to be necessary due to the style of performance encountered in this music, which at some point complicates the differentiation between onsets and vibrato or other effects. For annotating new samples the procedure described in [33] was adapted: the first and the third author of the paper did the annotations, while the fourth author corrected the results. Correcting the annotations means that it was only possible to delete annotations, and not to add new annotations. Each change in the correction process, except of a deletion, had to be discussed with the annotator. In this way cross-checked annotations were compiled for all the dataset. For the annotation the wavesurfer[2] software was used. Spectrogram, waveform and the F0 curves were used simultaneously to locate the onsets that were perceived in the sample.

Beside the data in MS as presented in Table I, 21 more samples of the instruments guitar, *ud*, piano, and violin were onset annotated. These files were used for parameter evaluations and development, and therefore this data set will be referred to as Development Set (DS) in the following sections. DS contains 674 onsets; see Table II for details. In the overall number of 78 samples that are contained in the main data set and in the development set, eight samples from the data set used in [33] are included, which contained an instrument listed in Table I (one file for cello, clarinet, piano, saxophone, trumpet, and violin, and two files for guitar).

Note that the focus of this paper lies on evaluating onset detection methods on monophonic pitched musical instrument sounds. Therefore, we recall that all the 78 collected samples of MS an DS contain only one instrument each. The data sets used in [33] and in [1] contain samples with several instruments playing together, which is referred to as complex mixture in [1]. In order to get a broader perspective, the complex mixture samples from these two publications were combined to a data set of thirteen complex mixture samples with an overall number of 498 onsets. This data will be referred to as complex mixture data set (CMS).

III. ONSET STRENGTH SIGNALS

As detailed above, it is the goal of this paper to evaluate three characteristics of musical instrument signals for their efficiency in onset detection: phase spectra (in terms of the phase slope function), magnitude spectra, and fundamental frequency contour. For this, the audio waveforms at a sampling frequency of 44.1 kHz are used to derive onset strength signals (OSS), which are expected to have local maxima at the samples which are related to musical onsets in the waveform. These OSS are computed using the sampling frequency of $f_{\mathrm{ons}} = 175$ Hz (5.7 ms). This sampling frequency guarantees a temporal solution which is equal to the minimal distance at which two sound events can be perceived separately, which was found to be at most 10 ms in [34].

A. Phase Slope

A signal $x[n]$ can be described in frequency domain by its Fourier transform $X(\omega) = A(\omega)e^{j\phi(\omega)}$, with ω denoting frequency and $A(\omega)$ being the amplitude spectrum and $\phi(\omega)$ being the phase spectrum. The basic motivation for using the phase

[2]http://www.speech.kth.se/wavesurfer/

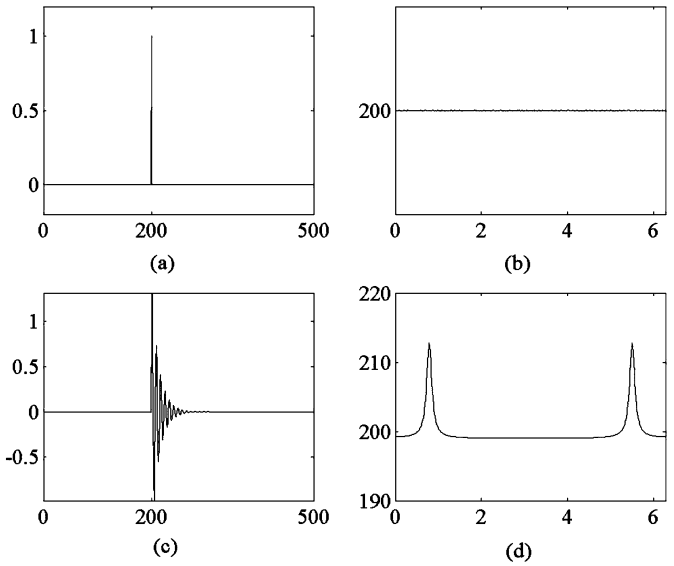

Fig. 1. (a) Unitary sample sequence delayed by 200 samples. (b) The group delay function of the signal in (a). (c) A minimum phase signal with an oscillation at $\pi/4$. (d) The group delay function of the signal in (c). (a) Sample, (b) Frequency (rad), (c) Sample, (d) Frequency (rad).

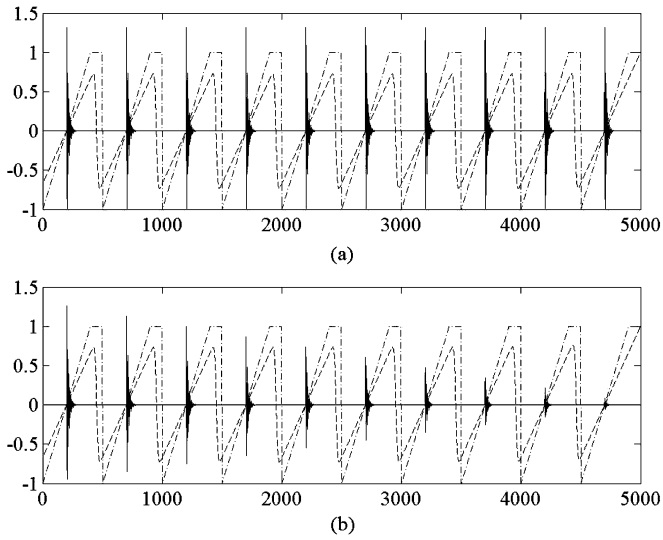

Fig. 2. (a) Sequence of impulses of constant amplitude and the associated phase slope function using long (dashed line) and short (dash-dotted line) window. (b) A sequence of impulses with linearly time varying amplitudes and the associated phase slope function using long (dashed line) and short (dash-dotted line) window. (a) Sample (b) Sample.

spectrum $\phi(\omega)$ of a signal for onset detection arises from properties of the group delay which is defined as

$$\tau(\omega) = -\frac{d\phi(\omega)}{d\omega} \tag{1}$$

The group delay of a delayed unit sample sequence $x[n] = \delta[n - n_0]$ is $\tau(\omega) = n_0$, $\forall\omega$. This holds because $x[n]$ has the Fourier Transform $X(\omega) = e^{-j\omega n_0}$ with the phase component $\phi(\omega) = -\omega n_0$. Computing its derivative regarding frequency (i.e., the group delay) results in $\tau(\omega) = n_0$, $\forall\omega$. This means that computing the average value of the group delay results in a value equal to the temporal distance between the center of the analysis window (at zero) and the position of the impulse (at n_0). This holds in general for the output of a minimum phase system excited by a delayed unit sample sequence as shown in [22]. Two simple examples are depicted in Fig. 1. The upper two panels (a) and (b) show the delayed unit sample sequence and its group delay, respectively. In the lower two panels in Fig. 1, the sequence shown in panel (a) is convolved with a minimum phase system, resulting in the signal shown in (c) and the corresponding group delay shown in (d). Note that in (d), the average of the group delay is again equal to the displacement between analysis window and the delay of the unit sample sequence. The peaks that appear in (d) are caused by the poles of the minimum phase system. Computing the average group delay, the influence of these poles disappears. This basic observation leads to the assumption that the onset of a note played by an instrument can be determined using group delay, because an instrument can be sometimes considered as a minimum phase system excited by an impulse. This impulse can be caused by, for example, a hammer, a bow, or the finger of a guitar player. An exception to this model is for example a violin player changing the left hand position while not changing the excitation, i.e., the movement of the bow. It is important to note that in [24] it has been shown that impulses can be detected with little impact of their actual amplitude by using group delay. Furthermore, it has been shown in [26] and [24], that onset (click) detectors based on the group delay are robust to additive noise as well. This means that even onsets that cannot be observed at all in magnitude can be detected using group delay.

In order to get a meaningful descriptor for onset strength from group delay as depicted in (1), the negative of its average is determined at each position of the analysis window. This value corresponds to the negative of the slope of the phase spectrum of the examined signal, and will thus be referred to as the phase slope $\tilde{\tau}$. In Fig. 2, the dashed lines show sequences of phase slopes obtained when shifting analysis windows over the depicted signals. It can be observed that at all points where the center of the analysis window coincides with the position of an impulse, the phase slope has a positive zero crossing. In Fig. 2, changing the length of the analysis window from signal period (long window) to a length shorter than the signal period (short window) does not affect this property of the phase slope. Also, as it was mentioned before, the efficiency of the phase slope function is not affected by the amplitude of the onset as it is clearly shown in 2(b). Thus, the fundamental idea is to detect the onsets of musical instruments by determining the positions of the positive zero crossings of the phase slope. Details on the computation of the slope function can be found in [24].

As observed already in [26], group delay cannot be used in such a straightforward way when dealing with music signals. Specifically, it was found in [26] that the group delay had to be computed in several frequency bands separately, while a selection of zero crossings was necessary. In this paper, parameters like the number of bands or criteria for the zero crossing selection are determined using only the development data set DS described in Table II.

The block diagram in Fig. 3 shows the processing steps for the computation of the phase slope onset strength signal (PS_OSS).

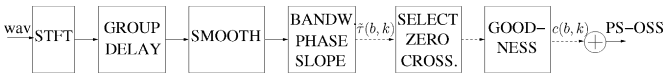

Fig. 3. Block diagram of the PS_OSS computation.

The first processing block consists of the computation of the short-time Fourier transform (STFT) of the signal $x[n]$

$$X(\omega,k)=\sum_{m=0}^{N-1}x[m+kh]w[m]e^{-j\omega m} \tag{2}$$

where hop size h is set to 5.6 ms in order to achieve the sampling frequency $f_{\text{ons}}=175$ Hz. The window length N of the applied Hanning window $w[n]$ has been set to $0.1s$. In order to apply the FFT algorithm the signal is zero padded. Note that in the context of the beat tracking task presented in [26], the analysis window length had been set to $0.2s$. It was found that reducing the window size to $0.1s$ leads to an improved Recall rate, which means that more of the annotated onsets get detected, while the number of false positives slightly increased. The second processing block contains the computation of group delays. To avoid the problems of unwrapping the phase spectrum of the signal for the computation of group delay, it is computed as in [35]

$$\tau(\omega,k)=\frac{X_R(\omega,k)Y_R(\omega,k)+X_I(\omega,k)Y_I(\omega,k)}{|X(\omega,k)|^2} \tag{3}$$

where

$$X(\omega,k)=aX_R(\omega,k)+jX_I(\omega,k)$$
$$Y(\omega,k)=Y_R(\omega,k)+jY_I(\omega,k)$$

are, respectively, the STFT of $x[n]$ and $nx[n]$ in analysis frame k, respectively.

In the third processing block, each frequency bin in the group delay vector $\tau(\omega,k)$ is median-filtered in time: $\tau(\omega,k)=\mu_{1/2}(\tau(\omega,k-i))$ for $i=[-4,-3,\ldots,4]$. This ninth-order median filtering is necessary due to the presence of many instruments with soft onsets in the dataset. It has been observed that especially for bowed string instruments onsets have a temporal extent of up to 50 ms, which is about the length of the median filter at the sampling period of 5.6 ms. Thus, this value represents an upper bound for the precision achievable on this data set. Next, the group delay vectors are divided into 21 non-overlapping frequency bands, as proposed in [8]. This transition to bandwise processing is indicated by dashed lines in Fig. 3. In each band, the negative of the median of the group delay values is determined, resulting in $[b=1\cdots 21]$ phase slope values $\tilde{\tau}(b,k)$ for each frame k. The exact number of bands was found to be uncritical, if this is chosen to be bigger than 5. Also, dividing the bands as proposed in [8] leads to a linear division for low frequencies. This was found to be crucial, because choosing for example logarithmic frequency bands causes the group delays of the lower bands to contain too few coefficients, and the medians are too noisy in these bands.

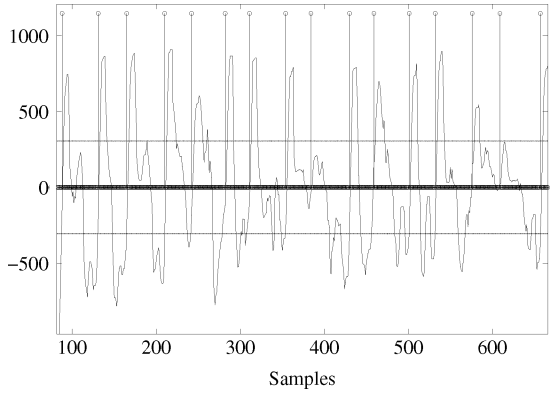

Fig. 4. Phase slope computed from the third band of a guitar signal.

As mentioned in [26], in each band the selection of zero crossings is necessary. In Fig. 4, the phase slope computed in the third band of a guitar signal is shown, along with the manually annotated onsets depicted as impulses. It can be observed that the phase slope has some spurious zero crossings, for example short after sample 100. It was observed that accepting only the positive zero crossings that are surrounded by large oscillations improves the accuracy of the detection. Such oscillations can easily be detected by thresholding, as shown by the dotted lines in Fig. 4. The positive threshold was determined by the mean of the absolute values of the phase slope for a whole sample; the negative threshold was simply the negative of this value. A positive zero-crossing is selected if the minimum and the maximum amplitude of the phase slope function, before and after the zero-crossing, respectively, pass the corresponding thresholds. Note that this way in the example shown in Fig. 4, there is no false positive detection. There are two missed onsets (at samples 384 and 609), because the phase slope has no peaks before and after these zero-crossings that cross the dotted threshold line. Different values of the threshold and the usage of adaptive thresholding have been tried, but the presented method was found to be sufficient. The next processing block in Fig. 3 is the goodness computation: for each of the above selected positive zero crossings, a value is assigned that denotes how much confidence can be given that this zero crossing coincides with an onset. In the context of speech excitation in [22], a method was proposed that measures the deviation of the computed phase slope from an ideal one (i.e., straight line). This approach was evaluated but a simpler solution was found: the confidence value is set to the value of the derivative of the phase slope in the vicinity of the zero crossing. High value of this derivative signifies high confidence. The output of this final block for the kth analysis window is the confidence level vector $c(b,k)$, that contains either the value zero in the bth band, when no zero crossing was selected, or the computed confidence value for this zero crossing. The final onset strength signal PS_OSS is then computed by summing $c(b,k)$ over the 21 bands: $\text{PS\_OSS}(k)=\sum_{b=1}^{21}c(b,k)$.

B. Spectral Flux

Spectral Flux (SF) is based on the detection of sudden positive energy changes in the signal which indicate attack parts of new

notes. Mainly there are two kinds of spectral flux OSS based on L1-norm and L2-norm as presented as follows:

$$\mathrm{SF\_OSS}_{L1}(k) = \sum_{\omega} H(|X(\omega,k)| - |X(\omega,k-1)|) \quad (4)$$

$$\mathrm{SF\_OSS}_{L2}(k) = \sum_{\omega} H(|X(\omega,k)| - |X(\omega,k-1)|)^2 \quad (5)$$

where $H(x) = (x+|x|)/(2)$ is the half-wave rectifier function, and $X(\omega,k)$ is the STFT of the signal with 5.6-ms hop size and a window length h of 46 ms. For the experiments in this paper, the L1-norm SF is used, because in [3] it was shown that L1-norm outperforms L2-norm. The accuracy of onset detection using SF_OSS and its computational simplicity were presented in [1], [3].

C. Fundamental Frequency Change

Considering the fact that note onsets are often difficult to observe in the amplitude in the case of pitched non-percussive instruments, it was decided to evaluate an additional onset strength signal. When playing for example a bowed string instrument, it is possible to create a new note onset by changing the position of the finger on the fingerboard while keeping the excitation caused by the bow constant. Because of the constant excitation, these onsets will be difficult to be observed in the phase slope as well. The only clear change is then the fundamental frequency (F0). Thus, it was decided to compute an OSS using the F0 estimations produced by the YIN algorithm [36].

At first, F0 estimations were calculated every t_{hop} ms. It is common practice to use the cent unit (obtained by the division of an octave into 1200 logarithmically equal partitions) for musical F0 analysis. Most of the musical pitch perception studies use this logarithmic measure of relative pitch which can be easily computed by

$$\mathrm{F0}_c = H(1200 \log_2(\mathrm{F0}_{\mathrm{Hz}}/c)) \quad (6)$$

where the reference frequency, the frequency of note lowest-C, is $c = 440 \cdot 2^{-69/12} \approx 8.1758$ Hz and $H(x)$ is again the half wave rectifier as introduced in (4). The application of the rectifier sets all values smaller than c to zero, including the points were the YIN estimator did not compute any pitch ($\mathrm{F0}_{\mathrm{Hz}} = 0$ Hz). The computed sequence of pitch values is checked at the points where no F0 has been computed by YIN (i.e., $\mathrm{F0}_c = 0$). This is either the case in silence parts, or at unstationary parts of the sound like instrument onsets. For this, a simple silence detector was applied. Whenever missing pitch values coincide with silence, the pitch was set to the pitch of the previous frames. Otherwise, the missing pitch values were set to the next computed pitch. This way, the robustness to silence parts and the accuracy of the onset estimation was improved. An example for this improvement is shown in Fig. 5: a typical example of the F0 estimation in the proximity of an onset for a cello signal. In this Figure, the F0 change at sample 245 is related to an onset, but the pitch estimator gives a correct F0 just after sample 260. In this example, samples 245 to 260 are non-silent frames and the pitch

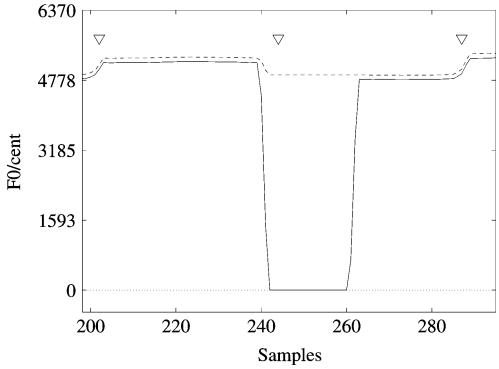

Fig. 5. F0 estimation for a cello sample before (bold) and after (dashed) silence filter, onsets at positions marked with arrows (samples 202, 245, and 288, respectively).

estimation is corrected, as shown by the dotted line, avoiding one false detection at sample 260. The final onset strength signal F0_OSS is computed from the silence-filtered F0 estimations F_0 as

$$\Delta F_0(t) = \min \begin{cases} \mathrm{mod}_{1200}(|F_0(t) - F_0(t-1)|)/600 \\ \mathrm{mod}_{1200}(-|F_0(t) - F_0(t-1)|)/600 \end{cases}. \quad (7)$$

This difference curve will have positives peaks at the instants of F0 changes. The magnitude of the peak depends on the change of F0. The modulo operator was applied to prevent from octave errors (note that 1200 cent is equivalent to an octave). The division by 600 normalizes the range of the F0_OSS from zero to one.

D. Fusion

It can be assumed that using each of the three OSS, it is possible to detect different types of onsets: SF captures onsets that are observable in magnitude change (hard onsets), F0 can detect note changes that happen in presence of a constant excitation, and PS can detect onsets that are characterized by the start of an excitation but that are not detectable in amplitude (soft onsets). As it is desirable not to select the optimal detector manually depending on the signal, a fusion of the three OSS can combine their advantages. In experiments combining the features to a three-dimensional space was tried (feature fusion), as well as the linear combination of the features to a single dimensional descriptor. Both approaches were not successful due to the following reasons: The sparseness of the OSS (many zero values) causes problems when trying to train classifiers in the three-dimensional space. Apart from that, the onsets are not exactly aligned in the three OSS: the beginning of an excitation is detected by PS while the maximum change in energy will be detected by SF, which happens typically with a temporal delay. A simpler solution that avoids these traps is the fusion at the decision level: Using each OSS separately onsets are determined. This results in three vectors: for each OSS, one decision vector with sampling frequency f_{ons} is obtained that has value one at the detected onsets and zero value elsewhere. By summing and smoothing these three vectors a fusion onset strength signal (FUSE_OSS) is obtained. Onsets can then be defined by a peak picking on FUSE_OSS. An example FUSE_OSS is shown

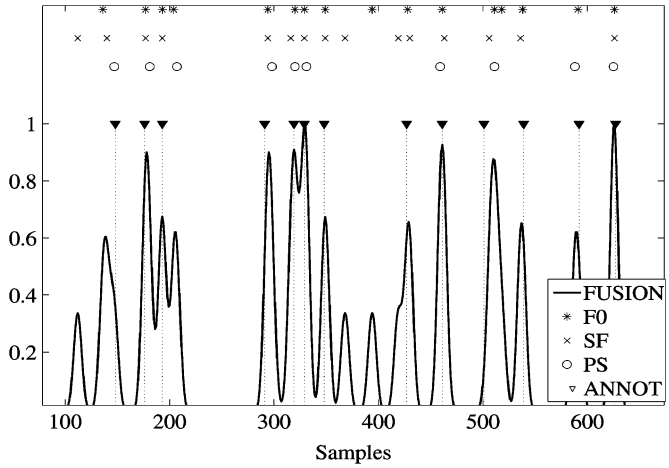

Fig. 6. Example for the decision fusion using a *ney* sample. Dotted lines with arrow markers show reference onset annotations, above the FUSE_OSS the positions of the onsets determined by F0_OSS, SF_OSS and PS_OSS are marked.

in Fig. 6. The dashed impulses show the reference onset annotations. The positions of the onsets determined by F0_OSS, SF_OSS and PS_OSS are also marked. It can be seen that the resulting FUSE_OSS has the largest maxima when all three OSS detect an onset close to this point. When only one OSS detects an onset this leads to a small amplitude in the FUSE_OSS, as for example at samples 370 and 390, where the onsets have only been detected by SF_OSS and F0_OSS, respectively. In the example shown in Fig. 6 there is an improvement by using the FUSE_OSS for onset detection. The general performance of FUSE_OSS will be provided in Section VI.

IV. Evaluation Methods

In the MIREX onset detection evaluation, the F-measure of the detection is computed as the main criterion. F-measure is defined as

$$F = \frac{2PR}{P+R} \tag{8}$$

with Precision, P, and Recall, R, being computed from the number of correctly detected onsets (N_{tp}), the number of false alarms (N_{fp}), and the number of missed onsets (N_{fn}) as

$$P = \frac{N_{tp}}{N_{tp}+N_{fp}}, \text{ and } R = \frac{N_{tp}}{N_{tp}+N_{fn}}. \tag{9}$$

According to the MIREX specifications, onsets are counted as correct detections when they are within a window of $t_{\text{tol}} = \pm 50$ ms around the onset annotation. If there are several onset detections in this tolerance window, only one is counted as true positive, the others are counted as false positives (double detections). If a detection is within the tolerance window of two annotations one true positive and one false negative are counted (merged onsets).

In order to get a more detailed description of the accuracy of an onset detector, the threshold δ applied to the OSS in the peak picking process (see Section V) can be varied in small steps. This way precision P and recall R values can be obtained for different threshold values, and P/R-curves are created by putting R values on the abscissa and P values on the ordinate. This representation has been proposed in the MIREX 2007 onset detection evaluation as well. In P/R-diagrams, the best onset detector, in terms of F-measure, is the one whose P/R-curve is closer to the upper right corner of the diagram. Furthermore, given a fixed threshold δ for the peak picking, the F-measures can be computed for varying time tolerances t_{tol}, to get an impression of how close the true detections are to the annotation in time. This gives a second representation besides the P/R-curves: plotting the F-measures over different tolerances in ms, which will be referred to as F/T-curve.

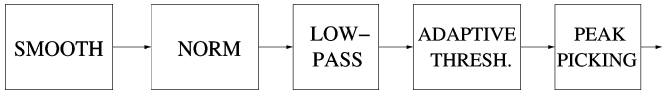

Fig. 7. Schematic of peak picking.

V. Peak Picking

In order to determine the time instants of onsets from the OSS described in Section III, the salient maxima in the OSS need to be detected. As mentioned in [1], this process is of major significance for the accuracy of the result. In Fig. 7 the basic blocks, as described in [1], of a peak picking process are depicted. In order to smoothen the onset strength signals they were filtered using a Hanning window of short length. Normalization refers to the subtraction of the means and the division by the variance of the OSS (z-score). The low pass filter is a simple third-order FIR filter with a cutoff frequency at $f_{\text{ons}}/5$. The adaptive threshold is computed by applying a moving median filter to the OSS. This threshold is then subtracted from the OSS to cancel dynamic changes. The length of the moving median filter was set to 17 samples (97.1 ms). The peak picking is a simple selection of local maxima. The performance of the OSS depends on the setting of a parameter δ, that defines the threshold that a local maximum has to excess in order to be selected as an onset. Threshold δ can be varied in small steps in order to create the P/R-curves described in Section II. An optimum threshold for an OSS can be obtained from the corresponding P/R-curve by determining which threshold leads to the best F-measure. Using this optimum threshold, the F/T-curves can be generated by changing the desired tolerance.

For each of the three OSS, the peak picking has been optimized by evaluating its accuracy according to the F-measure described in (8) on the development set. The resulting optimum peak picking procedure for the SF_OSS and F0_OSS are the same as described in [1], except of the additional computation of a smoothing (first block in Fig. 7) which is not mentioned in [2] as crucial; however, we found that this smoothing improved our results and therefore it was decided to include it. The length of the applied Hanning window was 51 ms. This degrades the possible resolution in time, but as detailed in Section II the required resolution is 100 ms because of the temporal extent of onsets of non-percussive instruments. Furthermore, the locations of the peaks can be preserved by applying zero-phase filtering.

For PS_OSS, the application of an adaptive threshold (fourth block in Fig. 7) computed using a moving median filter was

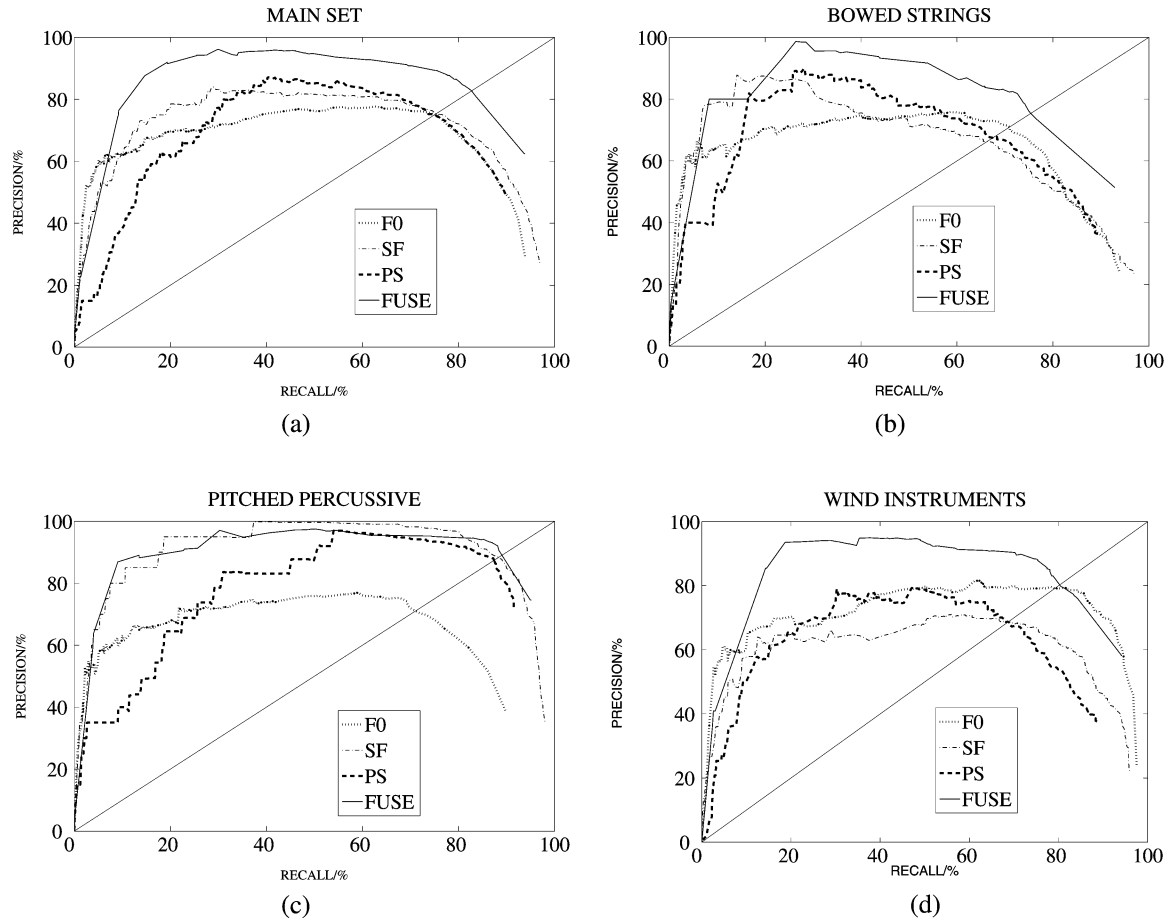

Fig. 8. Performance of the onset detection, Precision/Recall values in % are plotted on Ordinate/Abscissa, respectively. Curves have been obtained by changing threshold δ in small steps.

TABLE III
F-MEASURES OF THE OSS, ALONG WITH THRESHOLD VALUES δ

OSS	F0	SF	PS	FUSE
F-measure	74.1%	73.9%	73.7%	82.1%
δ	0.012	0.051	0.027	7.78

found to degrade the accuracy. This is due to the different characteristic of this OSS: it is not immediately derived from the temporal change of a signal property, but it is a time series of confidence values at some candidate onset instants, and contains more zero values than the other two OSS. Applying a moving median leads to the removal of too many onset candidates. Apart from that, the adaptive thresholds compensates changes in the strength signals due to changes in signal amplitude. These changes obviously do not affect the PS_OSS.

Thus, while for the peak picking in SF_OSS and F0_OSS all blocks of the diagram in Fig. 7 are active, for PS_OSS the fourth block (adaptive threshold) was left out.

VI. RESULTS

The performance of the OSS on the main data set is shown in the P/R-curves in Fig. 8. Fig. 8(a) shows the performance on the complete MS as described in Table I. Regarding their optimum F-measure, all three single OSS perform almost equally well, which can be seen from the fact that they cross the diagonal in the graph at almost the same point. The PS_OSS achieves higher precision values, while the SF_OSS is able to achieve higher recall rates. This means that when a low false alarm rate is of importance, as for example in beat tracking tasks, PS is superior to SF, which confirms the findings in [26]. Combining the decisions of the three OSS leads to a clear improvement of the performance. This can be seen from the large distance of the corresponding P/R-curve in Fig. 8(a), and from the best F-measures on the main set as listed in Table III. Here, the F-measure of the decision fusion (82.1%) improves the best single OSS F-measure (74.1%) by 8.0%. It is important to note that the three thresholds δ used in F0_OSS, SF_OSS and PS_OSS for the decision fusion are the ones that resulted in the best F-measure on the main data set MS. No significant difference was observed when these threshold values have been derived from either DS or the data from [1]. The three threshold values were left unchanged in the experiments conducted on the various subsets of the data (wind instruments, Turkish instruments, etc.) and on other data sets, in order not to present overoptimistic results for FUSE_OSS.

All the plots in Fig. 8 were produced using a ± 50 ms tolerance window. In Fig. 9, the F-measures are shown as a function of the tolerance value t_{tol}. This plot was generated using the MS data set, and the threshold values that produced the optimum F-measures as listed in Table III. It can be seen that for

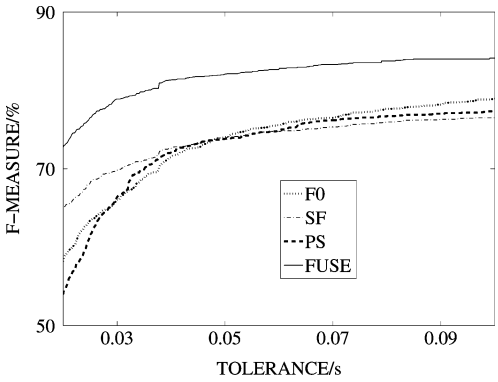

Fig. 9. F/T-curve for the three OSS and the decision fusion on the MS.

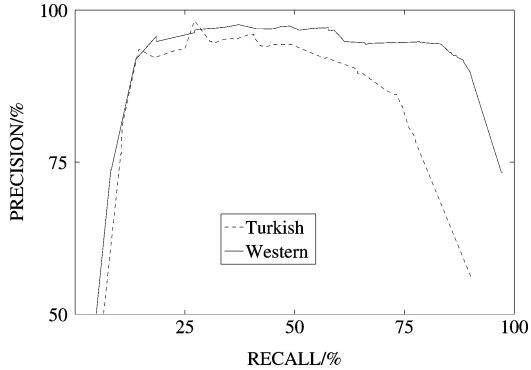

Fig. 10. Precision/Recall curves for the decision fusion, using western and Turkish performance style separately.

bigger tolerances, PS and F0 are superior to SF, but their performance decreases when demanding higher accuracy in time. For PS_OSS, this is due to the usage of the time median filtering in the phase slope computation, as described in Section III-A. The accuracy of PS_OSS can be improved by using a shorter median filter, which is possible when only hard onsets are considered. The decision fusion results in F-measures that are clearly superior for all desired tolerance values. The decreasing F-measures for F0_OSS for low tolerance values is caused by the uncertainty of the pitch estimation close to the onsets. However, when considering subfigures (b) and (d) in Fig. 8 the advantages of using a fundamental frequency estimation for onset detection can be observed: for both wind and bowed string instruments, F0_OSS achieves clearly improved F-measures compared to SF_OSS and PS_OSS; It is characterized by a curve that is closest to the upper right corner in both cases. Furthermore, for both wind and bowed string instruments the decision fusion can improve onset detection. The best F-measures of FUSE_OSS are higher than the best F-measures achieved with any single OSS. Moreover, FUSE_OSS improves the maximum precision. Note that for both instrument groups, using FUSE_OSS best precision values of more than 90% are reached. This leads to a very low false alarm rate when missing a number of onsets is accepted, which is typically desired in a beat tracking task as in [26]. For percussive onsets a well-known finding is confirmed: Because these types of onsets can be captured well from the magnitude spectrum, SF_OSS performs very well. Nevertheless, including also the information of the other OSS in the decision fusion leads to further improvement of the F-measure. The marginality of the improvement can be assigned to the bad performance of F0_OSS on this type of instruments. It was observed that the Yin F0 estimator had problems on these types of instruments, which are characterized by estimation errors especially in the vicinity of onsets.

As the data set presented in this paper contains some western instruments and some Turkish instruments, experiments could be conducted to judge the influence of the style of performance on the onset detection. For this, a set of instruments was chosen which contains only western performance styles (clarinet, guitar, piano, trumpet). The Turkish performance style is represented by the instruments: *kemençe* (bowed string), *ney* (wind instrument), *ud* and *tanbur* (both plucked string instruments). Note that both groups contain two instrument types with percussive and two types with soft onsets. Thus, the influence of the instrument types in this comparison is small, as the differences affect only the instrument timbre and not the type of onset. Decision fusion produced clearly superior results for both western and Turkish performance style. The resulting P/R-curves for the decision fusion are shown in Fig. 10. It can be seen that onset detection on the Turkish instruments is much more difficult. While similar maximum precision values can be achieved, the curve decreases rapidly when lowering the threshold of the detection (i.e., moving to higher recall rates). This coincides with an observation made in the onset annotation process: Turkish playing consists of many ornamentations which are difficult to annotate. These less salient onsets appear in the lower amplitudes of the OSS, and lead to the falloff of the P/R-curve. This form of the curve causes a decrease of the F-measure from 89.8% for the western instruments to 77.8% for the Turkish instruments. This shows that not only the type of instrument but also the style of performance affects the accuracy of an onset detection system. However, in order to specify exactly how much of this decrease is caused by playing style, comparative studies with the same instruments played in both styles must be conducted. This decrease in the performance is likely to be encountered in other improvised forms of music as well, such as the folk tunes investigated in [27].

For the complex mixture files in the CMS set described in Section II F0_OSS completely failed. This has to be expected since the Yin algorithm has been developed for F0 estimation from single sources. When complex mixtures are considered, an OSS will have to be derived from a multiple F0 estimator like the one described in [37]. On CMS, the PS and the SF onset strength signals were compared. Results show that about the same accuracy is obtained by using either of the two OSS. The obtained best F-measures on the data were 78.3% for the SF and 77.6% for the PS onset strength signals. The computed P/R-curves did not differ significantly. A decision fusion of only those two OSS resulted in a small improvement to an F-measure of 80.1%. Using a decision fusion of all three OSS for the CMS data results in an F-measure of 78.3%, i.e., the same F-measure as for SF_OSS alone. This shows that the proposed fusion method is robust even if one OSS completely fails. The influence of applying all subsets of OSS for decision fusion was evaluated on MS and all separate instrument groups, and the results are

TABLE IV
F-MEASURES WHEN USING ALL COMBINATIONS OF OSS FOR DECISION FUSION

	MS	P.PERC	WIND	BOWED	CMS
ALL	82.1	90.1	80.2	76.3	78.3
SF+PS	76.0	88.4	70.1	66.6	80.3
F0+PS	75.8	83.6	74.8	68.4	70.5
F0+SF	76.5	84.2	74.8	69.0	69.7

shown in Table IV. CMS data represents the only case in which decision fusion using all three OSS does not improve the detection performance. For all monophonic signals using all OSS leads to the best F-measures.

The data set presented in [1] was used for experiments as well. As this data set contains only 23 samples, it is difficult to use it for comparison with the results obtained on single instruments in this paper. This is because out of the 23 samples only one is a bowed string instrument, while wind instruments are not contained in this data set. It was decided to determine the average performance using all samples, and to exclude F0_OSS in these experiments. This is because there are seven complex mixture files and six non-pitched instrument files, and thus the usage of F0 on this data would be meaningless. On this data PS_OSS was slightly superior to SF_OSS in the sense of F-measure (90.4% compared to 89.0%). A fusion of these two OSS was not found to further improve results on this data.

VII. CONCLUSION

In this paper a novel phase slope based onset strength signals (PS_OSS) was introduced. PS_OSS is able to reach good performance when considering soft onsets, while high precision values can be reached using this descriptor. The proposed F0_OSS performs very well for soft onsets in the sense of F-measure, but has problems for hard onsets due to inaccuracies of the F0 estimator. Because of that, and in order to use F0_OSS on complex mixtures as well, an appropriate F0 estimator must be used. The decision fusion of the onset detections derived from SF_OSS, F0_OSS and PS_OSS was shown to improve independently from the type of signal. Thus, it constitutes a method to detect onsets from pitched musical instruments without the necessity of choosing any signal dependent parameters.

Considering the data set, it can be concluded that a diverse data set of pitched instruments is now available for the evaluation of onset detection systems. Requests can be addressed to the first author. Comparing the best F-measures of the presented data set using single OSS (73.3%) with the best F-measure of 90.4% achieved with a single OSS on the data set presented in [1] it can be concluded that the data set presented in this paper is more difficult and we expect it to be a valuable tool for researchers working in this area.

REFERENCES

[1] 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, no. 5, pp. 1035–1047, Sep. 2005.

[2] *Audio Onset Detection*, MIREX07 [Online]. Available: http://www.music-ir.org/mirex/2007/index.php/Audio_Onset_Detection

[3] S. Dixon, "Onset detection revisited," in *Proc. 9th Int. Conf. Digital Audio Effects*, 2006, pp. 133–137.

[4] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. New Music Res.*, vol. 36, no. 1, pp. 51–60, 2007.

[5] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME2000)*, New York, 2000, pp. 452–455.

[6] M. Goto and Y. Muraoka, "Beat tracking based on multiple-agent architecture—A real-time beat tracking system for audio signals," in *Proc. 2nd Int. Conf. Multiagent Syst.*, Kyoto, Japan, 1996, pp. 103–110.

[7] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 588–601, 1998.

[8] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP*, Washington, DC, 1999, pp. 3089–3092.

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

[10] M. Gainza, E. Coyle, and B. Lawyor, "Onset detection using comb filters," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New York, 2005, pp. 263–266.

[11] N. Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *Proc. 118th AES Conv.*, Barcelona, Spain, 2005, pp. 28–31.

[12] A. W. Schloss, "On the Automatic Transcription of Percussive Music From Acoustic Signal to High-Level Analysis," Ph.D. dissertation, Dept. Hear. Speech, Stanford Univ., Stanford, CA, 1985.

[13] B. Moore, B. Glasberg, and T. Bear, "A model for the prediction of thresholds, loudness and partial loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–239, 1997.

[14] S. Hainsworth and M. Macleod, "Onset detection in musical audio signals," in *Proc. Int. Comput. Music Conf. (ICMC)*, Singapore, 2003, pp. 163–166.

[15] J. P. Bello and M. Sandler, "Phase-based note onset detection for music signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP*, Hong-Kong, 2003, pp. 444–447.

[16] J. P. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *IEEE Signal Process. Lett.*, vol. 11, no. 6, pp. 553–556, Jun. 2004.

[17] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech Commun.*, vol. 49, no. 3, pp. 159–176, 2007.

[18] B. Yegnanarayana, D. K. Saikia, and T. R. Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-32, no. 3, pp. 610–623, Jun. 1984.

[19] H. A. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," *Signal Process.*, vol. 22, no. 3, pp. 259–267, 1991.

[20] K. Steiglitz and B. Dickinson, "Phase unwrapping by factorization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 6, pp. 723–726, Dec. 1982.

[21] R. M. Hegde, H. A. Murthy, and V. R. Gadde, "The modified group delay feature: A new spectral representation of speech," in *Proc. Interspeech-ICSLP*, , Korea, 2004, pp. 913–916.

[22] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 325–333, Sep. 1995.

[23] M. Brookes, P. A. Naylor, and J. Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 456–466, Mar. 2006.

[24] V. Kandia and Y. Stylianou, "Detection of clicks based on group delay," *Canadian Acoust.*, vol. 36, no. 1, pp. 48–54, 2008.

[25] A. Lacoste and D. Eck, "A supervised classification algorithm for note onset detection," *EURASIP J. Appl. Signal Process.*, vol. 2007, pp. 153–165, 2007.

[26] A. Holzapfel and Y. Stylianou, "Beat tracking using group delay based onset detection," in *Proc. of ISMIR—Int. Conf. Music Inf. Retrieval*, 2008, pp. 653–658.

[27] N. Collins, "Using a pitch detector for onset detection," in *Proc. ISMIR—Int. Conf. Music Inf. Retrieval*, Barcelona, Spain, 2005, pp. 100–106.

[28] J. P. Bello, C. Duxbury, M. E. Davies, and M. B. Sandler, "A combined phase and amplitude based approach for onset detection for audio segmentation," in *Proc. WIAMIS*, London, U.K., 2003, pp. 6–12.

[29] R. Zhou and J. D. Reiss, "Music onset detection combining energy-based and pitch-based approaches," in *Proc. MIREX Onset Detection Contest in ISMIR 2007—8th Int. Conf. Music Inf. Retrieval*, Vienna, Austria, 2007.

[30] C. C. Toh, B. Zhang, and Y. Wang, "Multiple-feature fusion based onset detection for solo singing voice," in *Proc. ISMIR—Int. Conf. Music Inf. Retrieval*, Philadelphia, PA, 2008, pp. 515–520.
[31] A. Kelleher, D. Fitzgerald, M. Gainza, E. Coyle, and B. Lawlor, "Onset detection, music transcription and ornament detection for the traditional irish fiddle," in *Proc. 118th AES Conv.*, Barcelona, Spain, 2005.
[32] M. Gainza, "Single note ornamentation transcription for the Irish tin whistle based on onset detection," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Naples, Italy, 2004.
[33] L. Daudet, G. Richard, and P. Leveau, "Methodology and tools for the evaluation of automatic onset detection algorithms in music," in *Proc. ISMIR—Int. Conf. Music Inf. Retrieval*, 2004, pp. 72–77.
[34] B. C. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. New York: Academic, Apr. 2003.
[35] A. Oppenheim, R. Schafer, and J. Buck, *Discrete-Time Signal Processing*. : Prentice-Hall, 1998.
[36] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, 2002.
[37] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc. ISMIR—Int. Conf. Music Inf. Retrieval*, 2006, pp. 216–221.

Andre Holzapfel received the graduate engineer degree in media technology from the University of Applied Sciences, Duesseldorf, Germany, and the M.Sc. and Ph.D. degrees in computer science from University of Crete, Heraklion, Crete, Greece.

He is currently working as a lecturer at the Technical Educational Institute of Crete. His research interests are in the field of speech processing, music information retrieval and ethnomusicology.

Yannis Stylianou (S'92–M'95) received the Diploma degree in electrical engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1991 and the M.Sc. and Ph.D. degrees in signal processing from the Ecole National Superieure des Telecommunications, ENST, Paris, France, in 1992 and 1996, respectively.

He is currently an Associate Professor in the Department of Computer Science, University of Crete, CSD UOC and an Associate Researcher in the Networks and Telecommunications Laboratory of the Institute of Computer Science ICS at FORTH. From 1996 until 2001, he was with AT&T Labs Research, Murray Hill and Florham Park, NJ, as a Senior Technical Staff Member. In 2001, he joined Bell-Labs Lucent Technologies, Murray Hill, (now Alcatel-Lucent). Since 2002, he has been with the Computer Science Department, University of Crete, and the Institute of Computer Science at FORTH, Heraklion, Crete, Greece. He holds nine patents. His current research focuses on speech signal processing algorithms for speech analysis, statistical signal processing (detection and estimation), and time-series analysis/modeling. He enjoys to work with speech and voice signals, music, and sounds produced by marine mammals.

Dr. Stylianou is on the Board of the International Speech Communication Association (ISCA), member of the IEEE Speech and Language Technical Committee and of the IEEE Multimedia Communications Technical Committee, on the Editorial Board of *Journal of Electrical and Computer Engineering*, JECE, Associate Editor of the EURASIP *Journal on Speech, Audio, and Music Processing*, ASMP, and Associate Editor of the EURASIP *Research Letters in Signal Processing*, RLSP. He is Vice-Chairman of the Cost Action 2103: "Advanced Voice Function Assessment," VOICE. He was Associate Editor for the IEEE SIGNAL PROCESSING LETTERS and on the Management Committee for the COST Action 277: "Nonlinear Speech Processing." Among other projects in FP6, he participated in the SIMILAR Network of Excellence coordinating the task on the fusion of speech and handwriting modalities. He is a member of ISCA and the Technical Chamber of Greece, TEE.

Ali C. Gedik received the B.Sc. degree in electrical engineering from Hacettepe University, Ankara, Turkey, in 1996 and the M.Sc. degree in musicology from Dokuz Eylul University, Izmir, Turkey. He is currently pursuing the Ph.D. degree in electrical engineering at the Izmir Institute of Technology, Izmir.

His research interests include the application of signal processing and machine learning techniques within the framework of computational ethnomusicology. He is the founder and administrative editor of the *Journal of Interdisciplinary Music Studies* (JIMS).

Barış Bozkurt received the Electrical Engineering degree and the Masters of Science degree in biomedical engineering both from Boğaziçi University, Istanbul, Turkey, in 1997 and 2000, respectively, and the Ph.D. degree in electrical engineering (speech processing) from Faculte Polytechnique De Mons, Mons, Belgium, in 2005.

After the Ph.D. degree, he was a Researcher with Svox AG, Zurich. He is currently an Assistant Professor with Izmir Institute of Technology (IYTE), Izmir, Turkey, where he continues his research on speech and audio signal processing.