Published version: IEEE Transactions on Audio, Speech, and Language Processing, 18(8):1968-1977, Nov. 2010. doi: 10.1109/TASL.2010.2040785

Auditory spectrum-based pitched instrument onset detection

Emmanouil Benetos, *Student Member, IEEE* and Yannis Stylianou, *Member, IEEE*

***Abstract*—In this paper, a method for onset detection of music signals using auditory spectra is proposed. The auditory spectrogram provides a time-frequency representation that employs a sound processing model resembling the human auditory system. Recent work on onset detection employs DFT-based features describing spectral energy and phase differences, as well as pitch-based features. These features are often combined for maximizing detection performance. Here, the spectral flux and phase slope features are derived in the auditory framework and a novel fundamental frequency estimation algorithm based on auditory spectra is introduced. An onset detection algorithm is proposed, which processes and combines the aforementioned features at the decision level. Experiments are conducted on a dataset covering 11 pitched instrument types, consisting of 1829 onsets in total. Results indicate that auditory representations outperform various state-of-the-art approaches, with the onset detection algorithm reaching an F-measure of 82.6%.**

***Index Terms*—Onset detection, group delay function, auditory spectrum.**

I. Introduction

THE detection of the starting time of each musical note plays an important role in the analysis of music signals. This process is referred to as musical instrument onset detection and it is an essential step for music transcription applications, as well as for music signal compression, beat tracking, audio editing applications, and music information retrieval. The goal of an onset detection system is the accurate estimation of note onset times, regardless of the instrument type or performance style.

Musical instruments can be roughly categorized into three families: non-pitched percussive, pitched percussive, and pitched non-percussive [1]. For the first category, the creation of an onset detection system is a relatively simple task, since percussive onsets are characterized by sudden energy changes. However, the detection of onsets of pitched instruments is a nontrivial task, since the way the onsets are produced is largely dependent on the instrument type. For a percussive instrument like the piano, onsets are located by amplitude changes in the spectrum. However, for a string instrument like the cello, the onset can be produced using a constant excitation, without any noticeable energy change. Several approaches for pitched instrument onset detection have been proposed in the literature [2], however they are mostly limited to a small number of instrument classes. In addition, most results presented in the literature concern individual instrument families, presenting overly optimistic performance rates. Most techniques rely on DFT-based features, neglecting psychoacoustic models that are able to mimic the human auditory system. It should be noted that there exists a difference between the actual onsets produced by instruments and perceptual onsets, caused by a noticeable change in intensity, pitch and timbre of the sound [3]. In [4], it was observed that onsets perceived by human annotators appeared late compared to automatic onset detection systems.

Most onset detection techniques focus on the creation of an onset detection function (also called onset strength signal or novelty function), whose peaks denote the presence of an onset. Onset detection functions are derived by detecting changes in certain audio signal features, and they usually have lower sampling rate compared to the original signal. Onsets are finally derived by employing a peak picking procedure on the onset detection function. Most onset detection approaches utilize features detecting changes in the energy or phase domain. Energy changes are most useful in detecting hard onsets, usually produced by percussive instruments, while phase-based changes are able to detect soft onsets, which denote the beginning of an excitation but may not indicate a change in signal energy [2], [5]. All proposed techniques in the literature employ processing in different frequency bands for the aforementioned features, which improves considerably onset detection accuracy. An example of an onset detection system using subband processing is the one proposed by Duxbury et al. [6], which uses subband decomposition and employs different descriptors for each frequency regions. Specifically for pitched instrument sounds, pitch detection techniques are also employed, which are able to detect onsets which are produced in the presence of constant excitation [5], thus not being detectable in the energy or phase domain. The major drawback of pitch-based onset detection techniques is that a change in pitch corresponds to the beginning of a stable time segment and not the actual onset time [7].

Onset detection systems related to this work will be addressed in detail. In [1], an onset detection system combining both energy and phase information was proposed. The generated detection function combines spectral flux and

E. Benetos is with the Dept. of Electronic Engineering, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom. He was with the Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH), GR-70013 Heraklion, Crete, Greece, and also with the Computer Science Department, Multimedia Informatics Lab, University of Crete, 71409 Heraklion, Greece. Y. Stylianou is with the Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH), GR-70013 Heraklion, Crete, Greece, and also with the Computer Science Department, Multimedia Informatics Lab, University of Crete, 71409 Heraklion, Greece (email: emmanouil.benetos@elec.qmul.ac.uk; yannis@csd.uoc.gr).

The authors would like to thank André Holzapfel, Ali C. Gedik, and Bariş Bozkurt for providing the onset annotated dataset.

This work was supported from the University of Crete Special Research Fund Account.

phase slope features in the complex domain. The employed dataset contained pitched nonpercussive, pitched percussive, nonpitched percussive, and complex sounds, comprising 1060 onsets in total. Reported results indicated an improvement over energy and phase-based approaches. An improved version of the system in [1] was proposed in [8], tested on the same dataset. The dataset used for evaluation in [8] was the same as in [1], with the inclusion of a separate dataset consisting of 106054 piano onsets derived from Mozart piano sonatas.

In [9], the negative derivative of the unwrapped phase over frequency, also called group delay function, was proposed for onset detection in a beat tracking application. The group delay function denotes the distance between the center of an analysis window and the position of an impulse created by a minimum phase signal. Since musical instruments can be considered as causal and stable filters, the usage of the group delay function was justified. Thus, onsets were detected by estimating the positions of positive zero-crossings of the group delay function. Multiband analysis was performed on two datasets, the first from the MIREX 2006 beat tracking task and the second containing samples of traditional Cretan music. In [5] the group delay function, the spectral difference and the fundamental frequency change features were combined at decision level for an onset detection system. A dataset containing samples from 11 different pitched instruments was developed, which is also used in the present experiments. The fundamental frequency was extracted using the YIN algorithm [10] and the aforementioned features were combined at decision level. Results indicated an F-measure of 82.1% for the fused onset detection function, which was an improvement of about 11% for each single descriptor.

In [4], a system for onset detection employing a constant-Q pitch detector was proposed, tested on the pitched non-percussive sounds also employed in [1]. The pitch detection algorithm employs a maximum likelihood approach, attempting to correlate a harmonic template with the constant-Q spectrum in order to find the best fitting fundamental frequency. An algorithm for vibrato suppression is also introduced, while results indicate an improvement over phase-based and energy-based approaches. The actual onset times were derived by the stable segment times by simply subtracting about 140ms from the detected time instant. It is also suggested in [4] that a detector based on a computational auditory model might improve onset detection performance. Another onset detection system employing pitch estimation was introduced in [7], which utilizes an energy descriptor for detecting hard onsets and a pitch descriptor for detecting soft onsets. Instead of the DFT, the resonator time-frequency image was used for feature extraction [11]. For the detection of the onset times from the stable segments derived by pitch, a window of length 300ms looking backward from the beginning from the stable segment was used, searching for increase in energy at the frequency bin corresponding to the detected pitch. The MIREX 2007 onset detection dataset was employed and the method proposed in [7] outperformed all competing approaches.

As far as systems employing psychoacoustically motivated models are concerned, a filterbank system was introduced in [12], which uses the loudness of each band as a descriptor. In [13], the perceptual spectral flux was introduced as a feature for onset detection, which employs a weighted STFT, weighting frequency bands according to the human loudness contour. In [14], a comparison between the approach in [1] and the psychoacoustically motivated approaches in [12], [13] was performed. Finally, Gainza et al. [15] employed FIR comb filters on a frame by frame basis combining the inharmonicity properties with the energy increases of the signal onset.

The auditory spectra, based on the model presented in [16], are designed to mimic the functions of the human auditory system. In this paper, an approach for onset detection is proposed by employing auditory spectra instead of DFT-derived spectra for the computation of onset detection features. More specifically, the group delay function and the spectral flux are derived in the auditory framework. In addition, an auditory spectrum-derived fundamental frequency estimator based on the harmonic product spectrum technique is developed. Then, an onset detection system is finally suggested, combining the aforementioned three descriptors at decision level. Comparative experiments on onset detection were performed using the complex domain phase and energy descriptor used in [2]. The employed dataset for experimentation was introduced in [5] and contains a wide variety of pitched instrument types, not limited to western instruments, containing 1829 onsets in total. Results indicate that the isolated auditory descriptors reach high precision values for onset detection, outperforming the respective DFT-based features. Finally, the combined onset strength signal of the three auditory spectrum-derived descriptors performs slightly better compared to state-of-the-art approaches for onset detection employing DFT-based techniques, reaching an F-measure of 82.6%.

The outline of the paper is as follows. The employed auditory model is presented in Section II. In Section III, the spectral flux and group delay function are derived for the auditory spectrum, a fundamental frequency estimation algorithm using the auditory model is presented and the proposed onset detection system is discussed. The employed dataset, the methods used for evaluation and the experimental results are discussed in Section IV. Conclusions are drawn and future directions are indicated in Section V.

II. Auditory Model

The auditory model was first introduced in [17] and formalized in [16]. It is inspired by physiological, psychoacoustical and computational studies in the human primary auditory cortex. The model consists of two stages, a spectral estimation model (designed to mimic the cochlea in the auditory system) and a spectral analysis model (which mimics the primary auditory cortex). The spectral estimation model produces the so-called auditory spectrogram.

The auditory spectrum produces a time-frequency representation of the signal on a logarithmically scaled frequency axis, referred as the tonotopic axis. The auditory spectrogram consists of 128 log-frequency bins and can be approximated as:

$$X_A[n,l] = \max(\partial_l g(\partial_n x[n] *_n h[n,l]), 0), \qquad (1)$$

where $x[n]$ is the original signal and $h[n,l]$ is a minimum-phase seed bandpass filter where $h[n,l] = \alpha h[\alpha n, l_0]$, with scaling factor $\alpha = 2^{l-l_0}$ and $l = 1,\ldots,129$. The Fourier transform of $h[n,l]$ for a given l satisfies the following property: $\frac{\partial H(\omega)}{\partial \omega} = j\omega H(\omega)$. The convolution of $x[n]$ with $h[n,l]$ is an application of a constant-Q filter-bank wavelet transform. In (1), ∂_i denotes the partial derivative over i, and $g(m) = \frac{1}{1+e^{-m}} - \frac{1}{2}$ is a sigmoid-like function, which is used to model the hair cell response in the human auditory system. It should be noted that in (1) two operations are not mentioned for purposes of simplicity. The first consists of a temporal smoothing operation which filters out responses beyond 4 kHz and the second consists of a temporal integration of $X_A[n,l]$, which is followed by subsampling. In general, the auditory spectrogram is relatively insensitive to broadband changes in the spectral shape and it is more robust against noise compared to the DFT-derived spectrogram [16].

III. Auditory Spectrum-based Onset Detection

In this section, the system developed for onset detection using auditory spectra will be presented. First, the phase slope and spectral flux features will be derived for the auditory framework. Then, an algorithm for fundamental frequency estimation using auditory spectra will be suggested. Finally, the fusion of the aforementioned descriptors will be discussed.

A. Auditory Group Delay

In [5], [9] the group delay function was computed in the DFT domain as:

$$GRD[\omega,k] = \frac{X_R[\omega,k]Y_R[\omega,k] + X_I[\omega,k]Y_I[\omega,k]}{|X[\omega,k]|^2} \quad (2)$$

where $X[\omega,k]$ and $Y[\omega,k]$ are the DFTs of $x[n]$ and $nx[n]$, respectively and the subscripts $_R$ and $_I$ denote the real and imaginary parts of the DFTs. The auditory spectrum however does not contain any imaginary parts, thus phase information will be extracted using the analytic signal representation of the auditory spectrum [18]. In our analysis, the phase slope (also called instantaneous frequency) in the auditory spectrum is computed in a way similar to [19]. For simplicity purposes, n will refer to the continuous domain. The following approximation for the auditory spectrum will be used:

$$X_A[n,k] \approx x[n] *_n h[n,l] \quad (3)$$

taking into account the constant-Q wavelet transform, which is the most important part in the auditory spectrum formulation. The analytic signal representation of $X_A[n,l]$ is defined as:

$$\mathcal{X}_A[n,l] = X_A[n,l] + j\hat{X}_A[n,l], \quad (4)$$

where $\hat{X}_A[n,l]$ is the Hilbert transform of $X_A[n,l]$ over n [20]. The analytic signal can also be expressed in polar form as:

$$\mathcal{X}_A[n,l] = \mu[n,l]e^{j\phi[n,l]}, \quad (5)$$

where $\mu[n,k]$ is the magnitude of $\mathcal{X}_A[n,l]$ and $\phi[n,l]$ its respective phase. Following (5), $\phi[n,l]$ can be expressed as:

$$\phi[n,l] = \Im(\ln(\mathcal{X}_A[n,l])) \quad (6)$$

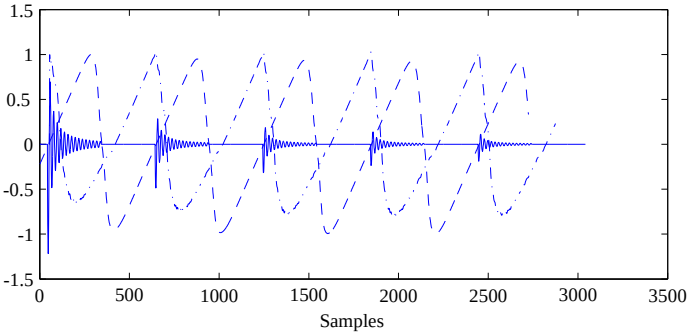

Fig. 1. A sequence of impulses with linearly time varying amplitudes, the associated DFT-based group delay function (dashed line), and the associated auditory spectrum-based group delay function (dashed-dotted line).

Thus, the negative derivative of the phase over n of the analytic signal is expressed in the form:

$$\tau[n,l] = -\frac{\partial \phi[n,l]}{\partial n} = -\Im\left(\frac{1}{\mathcal{X}_A[n,l]} \cdot \frac{\partial \mathcal{X}_A[n,l]}{\partial n}\right) \quad (7)$$

In the following analysis, the Fourier transforms of the auditory spectra will be utilized for convenience. Thus:

$$\mathcal{F}\{X_A[n,l]\} = X(\omega)H(\omega,l) \quad (8)$$

$$\mathcal{F}\left\{\frac{\partial X_A[n,l]}{\partial n}\right\} = j\omega X(\omega)H(\omega,l) \quad (9)$$

$$\mathcal{F}\left\{\frac{\partial \hat{X}_A[n,l]}{\partial n}\right\} = \text{sgn}(\omega)\omega X(\omega)H(\omega,l) \quad (10)$$

where $X(\omega)$ and $H(\omega,l)$ denote the Fourier transform of $x[n]$ and $h[n,l]$, respectively. We propose $Y_A[n,l] = x[n] *_n nh[n,l]$. By employing the definition of $h[n,l]$ in Section II, $Y_A[n,l]$ can be formed in the Fourier domain as:

$$\mathcal{F}\{Y_A[n,l]\} = -\omega X(\omega)H(\omega,l) \quad (11)$$

$$\mathcal{F}\{\hat{Y}_A[n,l]\} = j\text{sgn}(\omega)\omega X(\omega)H(\omega,l) \quad (12)$$

Using (9) and (12), it can be seen that:

$$\begin{aligned}\mathcal{F}\{\hat{Y}_A[n,l]\} &= \text{sgn}(\omega)\mathcal{F}\left\{\frac{\partial X_A[n,l]}{\partial n}\right\} \Rightarrow \\ \hat{Y}_A[n,l] &= \frac{1}{2}\left(\delta[n] + \frac{j}{\pi n}\right) *_n \frac{\partial X_A[n,l]}{\partial n} \\ &= \frac{1}{2}\frac{\partial X_A[n,l]}{\partial n} + \frac{1}{2}j\frac{\partial \hat{X}_A[n,l]}{\partial n} \end{aligned} \quad (13)$$

Likewise, using (10) and (11):

$$\begin{aligned}\mathcal{F}\{Y_A[n,l]\} &= -\text{sgn}(\omega)\mathcal{F}\left\{\frac{\partial \hat{X}_A[n,l]}{\partial n}\right\} \Rightarrow \\ Y_A[n,l] &= -\frac{1}{2}\left(\delta[n]) + \frac{j}{\pi n}\right) *_n \frac{\partial \hat{X}_A[n,l]}{\partial n} \\ &= -\frac{1}{2}\frac{\partial \hat{X}_A[n,l]}{\partial n} + \frac{1}{2}j\frac{\partial X_A[n,l]}{\partial n} \end{aligned} \quad (14)$$

Combining (13) and (14):

$$\begin{aligned}\mathcal{Y}_A[n,l] &= Y_A[n,l] + j\hat{Y}_A[n,l] \\ &= -\frac{\partial \hat{X}_A[n,l]}{\partial n} + j\frac{\partial X_A[n,k]}{\partial n} \\ &= j\frac{\partial \mathcal{X}_A[n,k]}{\partial n} \end{aligned} \quad (15)$$

Substituting (15) into (7):

$$\tau[n,l] = \Re\left(\frac{\mathcal{Y}_A[n,l]}{\mathcal{X}_A[n,l]}\right) = \frac{Y_A[n,l]}{X_A[n,l]} \quad (16)$$

It can be seen that the definition of $\tau[n,l]$ in (16) closely resembles the definition of the DFT-based group delay, by removing the imaginary parts of (2). In addition, it can be seen that no average value has been computed for neighboring bands as is the case for the DFT-based phase slope. Thus, the usage of the term *auditory group delay* instead of auditory phase slope is justified. In Fig. 1, an example of a DFT-based phase slope function is depicted by the dashed line which has positive zero crossings at the position of impulses in the signal, along with the auditory spectrum-based group delay function, depicted as a dashed-dotted line, which is obtained when shifting an analysis window over a sample signal. It can clearly be seen that the auditory group delay exhibits peaks at the position of impulses in the signal.

The processing steps for the computation of the onset detection signal based on the auditory group delay function, can be seen in Fig. 2. The auditory spectrum is computed using the NSL toolbox [21]. For the computation of the auditory spectrum the window length is set to 0.1s, with 4.5ms hop size and the resulted spectrogram is computed for a bandwidth of 76-3242 Hz. In processing block 2, the auditory group delay function is computed from auditory spectrograms $X_A[n,l]$ and $Y_A[n,l]$ using (16). For our analysis, tonotopic bands $b = 1,\ldots,120$ of the auditory spectrogram were utilized, thus ignoring bands containing high-frequency noise. In processing block 3, the group delay deviation over n is calculated for each band, by using two frames as a hop size, as in [1], [2]:

$$\Delta GRD[n,l] = \Delta\tau[n,l] + \Delta\tau[n-1,l] = \tau[n,l] - \tau[n-2,l] \quad (17)$$

where $\Delta\tau[n,l] = \tau[n,l] - \tau[n-1,l]$. In processing block 4 of Fig. 2, each band is smoothed in time using a 3rd degree Savitzky-Golay filter with window size equal to 12 samples [22]. The Savitzky-Golay filter uses local polynomial regression and is considered superior compared to FIR filters or moving average filters, preserving the local maxima of the signal while rejecting noise. In processing block 5, for each band of $\Delta GRD[n,l]$, peak picking is performed in order to select candidate onsets. For each band, an onset detection signal is constructed containing either the value zero when no peak has been detected, or the amplitude of the detected peak. In each band b, a threshold for peak detection is determined separately by the mean value of the half-wave rectified auditory group delay function for the particular band. Finally, all band-wise detection signals are summed, creating a single onset detection signal based on the auditory group delay. In Fig. 3, the band-wise mean of the smoothed ΔGRD for a tanbur (plucked string instrument) recording is depicted. By peak picking on the detection signal, seven onsets are detected while two missed detections occur.

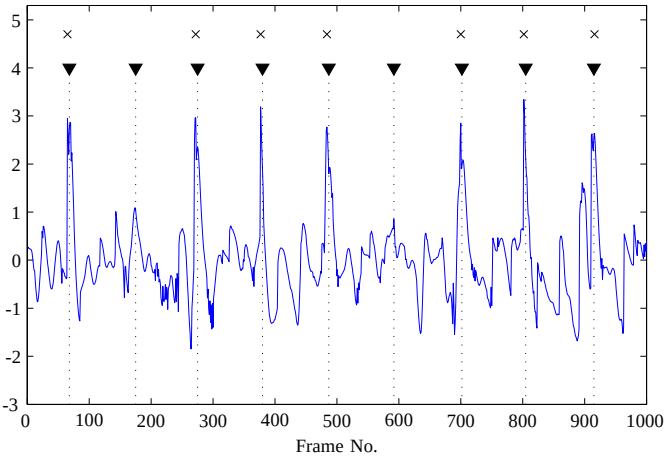

Fig. 3. The band-wise mean ΔGRD of a tanbur recording. The dotted arrows correspond to the ground truth, while the 'x' marker correponds to onsets detected by peak picking on the detection signal.

Onsets are detected from the auditory group delay onset strength signal by a selection of local maxima. First, the auditory group delay detection function is normalized using z-score standardization. Afterwards, a moving median filter of length 0.2s is computed as an adaptive threshold, which is a robust method for detecting impulses in audio signals [23]. The adaptive threshold is then subtracted from the detection signals. Finally, peak picking is performed, by selecting peaks that are higher than threshold δ and are separated by a minimum peak distance of 40ms.

B. Auditory Spectral Flux

The spectral flux in the Fourier domain measures the magnitude changes in each frequency bin [8] which indicate attack parts of new notes [5]. The spectral flux can be used effectively for onset detection of percussive signals, but its performance decreases when soft onsets are located [1]. In [5], the spectral flux was computed using the L1 norm:

$$SF[k] = \sum_{\omega} HW(|X[\omega,k]| - |X[\omega,k-1]|) \quad (18)$$

where $X[n,k]$ is the DFT of the original signal at the k-th frequency bin and n-th frame and $HW(x) = \frac{x+|x|}{2}$ is the half-wave rectifier function. The descriptor combining phase and spectral flux proposed in [1], [8] which will be used for performance comparison results in Section IV is formulated as follows:

$$CD[n] = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |X[n,k] - |X[n-1,k]|e^{\psi[n-1,k]+\psi'[n-1,k]}| \quad (19)$$

where $\psi[n,k]$ is the phase of $X[n,k]$.

The spectral flux in the auditory domain is defined in a similar manner to the DFT-based spectral flux in (18), using

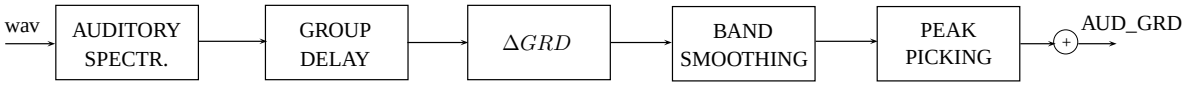

Fig. 2. Block diagram of the computation of the auditory spectrum-based group delay.

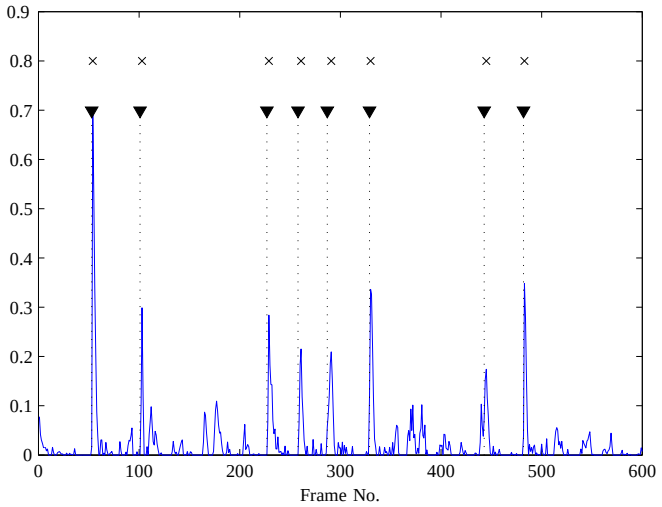

Fig. 4. The auditory spectral flux of a trumpet recording. The dotted arrows correspond to the ground truth, while the 'x' marker correponds to onsets detected by peak picking on the auditory spectral flux.

the L1 norm:

$$AUD\_SF[n] = \sum_{l} HW(X_A[n,l] - X_A[n-1,l]). \quad (20)$$

It should be noted that the L1 norm is favored over the L2 norm for spectral flux computation [8]. For the creation of an onset strength signal derived from the auditory spectral flux, the original signal is resampled to 8kHz and the spectral flux is computed with a step size of 8*ms*, without overlapping. It should be noted that no band-wise smoothing or band selection was performed on the auditory spectral flux, since it was found to degrade onset detection performance. This was the case in soft onsets produced by string instruments, where the envelope of the spectral flux detection function was much smoother compared to the envelope of the detection function produced by hard onsets. Onsets from the auditory spectral flux are detected in a similar way to the auditory group delay onset strength signal: z-score normalization is performed, a moving median filter of 0.2s is employed as an adaptive threshold, and finally peak picking is performed using the threshold δ. In Fig. 4, the auditory spectral flux of a trumpet recording is depicted. As can be seen, onsets derived from the auditory spectral flux occur within a short delay compared to the actual annotation. This is due to the fact that the spectral flux detects the largest energy increase in the attack part of the note and not the actual onset time.

C. Fundamental Frequency Estimation

While hard onsets can be easily detected using algorithms measuring energy differences over time, soft onsets which are produced by pitched non-percussive instruments are harder to detect using the aforementioned methods, due to the fact that the sounds are produced with a constant excitation [5], [7]. Thus, the only detectable change of a soft onset is in the pitch domain. A brief overview for pitch detection techniques for music processing can be found in [24]. As far as onset detection experiments using pitch estimation are concerned, Collins [4] employed a constant-Q pitch detector, choosing a best matching harmonic template, which is a variant of the maximum likelihood (ML) approach [25]. Zhou et al. proposed a pitch detection algorithm based on the resonator time-frequency image [7]. Finally, Holzapfel et al. [5] utilized the YIN fundamental frequency estimator [10], which is a modified version of the autocorrelation method for pitch estimation.

For our experiments, we propose a fundamental frequency estimation algorithm which is based on auditory spectra. The algorithm is inspired by the harmonic product spectrum (HPS) algorithm, which was proposed by Noll [25]. The HPS algorithm is based on multiplying spectral frames which are subsampled by different integer values as to align the harmonics. The resulting product of the subsampled spectral frames is searched for a maximum value which denotes the fundamental frequency for the current frame. The HPS algorithm is relatively inexpensive due to FFT computation and resistant to noise, compared with ML approaches. A drawback of the HPS algorithm is the presence of octave errors in fundamental frequency measurements, which however can be corrected by post-processing operations [24].

Since the auditory spectrum produces 24 coefficients per octave (1 bin corresponds to 50 cent units), the first step for the proposed algorithm is the creation of a semitone-resolution auditory spectrogram:

$$P_A[n,m] = X_A[n,2m-1] + X_A[n,2m] \quad (21)$$

where $m = 1,\ldots,64$. Since the auditory spectrogram is defined on a logarithmic scale, the proposed auditory pitch spectrogram is defined as a product of translated auditory spectral frames:

$$Q_A[n,m] = P_A[n,m] \cdot P_A[n,m-12] \cdot P_A[n,m-24] \quad (22)$$

The estimated fundamental frequency over an auditory spectral frame corresponds to the semitone with the highest value in $Q_A[n,m]$:

$$F_0[n] = \max_m \{Q_A[n,m]\} \quad (23)$$

In Fig. 5, the auditory pitch spectrogram $Q_A[n,m]$ of a saxophone recording used in the present experiments is displayed. Darker color in the auditory pitch spectrogram corresponds to higher values, and each frame has a length of 8ms. The note sequence of the recording ($g_3 - e\flat_3 - g_3 - f_3 - e\flat_3 - f_3 - g_3 - e\flat_3 - c_3 - g_2$) is quite evident in the auditory

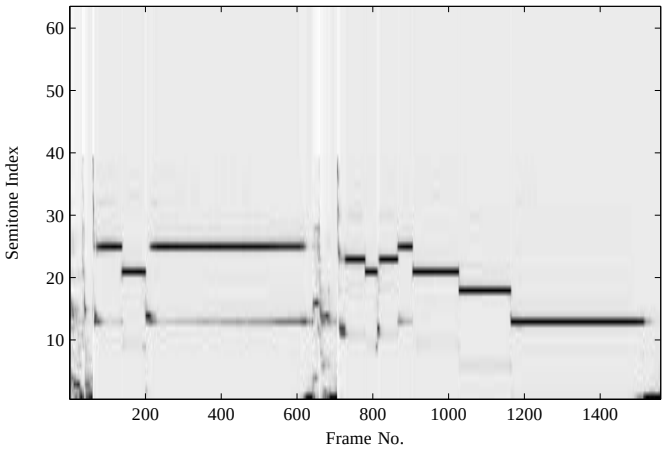

Fig. 5. The auditory pitch spectrogram $Q_A[n,m]$ of a saxophone recording.

pitch spectrogram. The presence of low frequency noise which can be easily removed via silence filtering can also be noticed in the beginning of the recording and around frames 600-700, as well as an octave error around frames 1500-1600.

In Fig. 6, the basic blocks for computing the fundamental frequency-based onset strength signal using auditory spectra are depicted. In order to compute the onset strength signal, the changes over the auditory spectrum-based fundamental frequency estimator will be taken into account. In our approach, a modified version of fundamental frequency changes used in [5] is employed:

$$\Delta \mathrm{F}_0[\mathrm{n}] = |\mathrm{mod}_{11}(\mathrm{mod}_{12}(F_0[n])) - \mathrm{mod}_{11}(\mathrm{mod}_{12}(F_0[n-1]))| \quad (24)$$

From (24) it can be seen that first octave errors are removed and afterwards spaces of 11 semitones are also considered as octave errors and subsequently removed. Smoothing on $\Delta F_0[n]$ is performed afterwards using a 3rd degree Savitzky-Golay filter on a 64ms window [22]. A simple silence detector is applied into the smoothed $\Delta F_0[n]$, using the short-time energy of the auditory pitch spectrum, $Q_A^2[n,m]$. Whenever a segment in $\Delta F_0[n]$ contains a fundamental frequency change whose value in the energy pitch spectrum is lower than a threshold, $\Delta F_0[n]$ is set to zero. Thus, it is also possible to detect onsets where the produced notes have the same pitch. Local maxima on the silence-filtered $\Delta F_0[n]$ are subsequently selected using a peak picking algorithm. The aforementioned peaks correspond to the starting points of the steady-state parts of musical notes. In order to estimate the actual onset times from the starting points of the stable segments, an approach similar to [7] was adopted. For each stable segment, a window of length 230ms is employed, looking backwards from the beginning of the stable segment. The energy of the semitone-resolution auditory spectrogram $P_A^2[n,m]$ is considered. The actual onset time is derived by peak picking on $P_A^2[n,m_0] - P_A^2[n-1,m_0]$, where m_0 is the spectral bin corresponding to the estimated fundamental frequency for the stable segment. Thus, onset times are estimated by searching for the largest energy increase in the bin corresponding to the estimated fundamental frequency. It should be mentioned that threshold δ of detecting changes in fundamental frequency in the 4th block is set to one semitone, but can be adjusted in order to create P/R curves, as will be explained in Section IV-B.

D. Fusion

It was observed in [1], [5], [7] that a combination of descriptors for onset detection leads to improved results, since different types of onsets are detected for each onset strength signal. The spectral flux detects hard onsets, which exhibit a sudden change in energy, while phase-based descriptors are able to detect softer onsets, regardless of the signal energy. The major drawback of spectral flux features is their inability to detect onsets of nonpercussive sounds. In addition, the major drawback of phase-based approaches is their susceptibility to noisy low-energy components and phase distortions due to post-production treatments [2]. Finally, pitch-based features are suitable for detecting soft onsets with a gradual change in energy, which are more difficult to detect in the energy and phase domain. Pitch-based descriptors are not however suitable for detecting percussive sounds which do not contain tonal components.

A system classifying each onset as hard or soft was proposed in [7], employing energy or pitch-based descriptors, respectively. However, this approach neglects onsets detectable by phase changes, which roughly lie between the two aforementioned categories. In addition, the combination of energy and phase-based descriptors at the complex domain as shown in (19) will be also addressed in comparative experiments in Section IV-C.

In our approach, we employ a fusion scheme similar to the one proposed in [5], combining the three descriptors at the decision level. This choice is justified by the fact that the onset strength signals are not aligned: the spectral flux detects the time instant with the maximum energy change at the attack part of the signal, the phase slope detects the instant marking the beginning of the excitation (the actual onset time), while the fundamental frequency detects the beginning of the stable part of the musical note, which in some cases may occur even 0.2s after the actual onset. Thus, from each of the three descriptors an onset strength signal is obtained, containing either the value one at the instant of the detected onset or zero otherwise. Since the spectral flux and fundamental frequency onset strength signals have the same sampling frequency, the auditory group delay-based signal is downsampled to match the step size of 8ms for the aforementioned signals. All three signals are summed and smoothed using a moving median filter of 40ms length, thus creating the fused onset strength signal. Peaks are detected being separated by a minimum peak distance of 40ms, thus avoiding multiple peak selection for one actual onset.

IV. Experiments

In this section, the experimental procedure used for onset detection will be addressed. The employed dataset will be described, the evaluation measures used for assessment will be defined, and the experimental results of the various descriptors, along with comparative experiments, will be presented.

Fig. 6. Block diagram of the computation of the fundamental frequency-based onset strength signal.

Instrument	No. of onsets	No. 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
Total	**1829**	**57**

TABLE I
ONSET DATASET DETAILS.

A. Dataset

For testing the performance of onset detection systems, an annotated dataset covering several musical instrument classes with a wide range of pitch and dynamics is essential [8]. Further discussion on the creation of a dataset for onset detection can be found in [26]. In our experiments, the dataset introduced in [5] was employed. It consists of 57 recordings of pitched instruments, covering 11 instrument types, as seen in Table I. A smaller dataset was employed for parameter tuning as in [5], which consisted of 21 guitar, piano, ud, and violin recordings, containing overall 674 onsets. The annotation has been done using the *wavesurfer* program [27], and has been validated by three authors of [5]. For the test dataset, the various instrument types can be organized into three classes: pitched-percussive instruments (guitar, ud, piano, and tanbur), wind instruments (clarinet, ney, saxophone, and trumpet), and bowed string instruments (cello, kemençe, and violin). It should be noted that the set is not limited to western instruments, but also contains middle-eastern instrument samples, thus providing a more comprehensive view on the effectiveness of onset detection features on various instrument types. Recordings from non-pitched percussive instruments are not included, since their onsets can be easily detected using energy or amplitude-based descriptors. This dataset is deemed more suitable for experiments in onset detection compared to other compiled sets, since each instrument type contains roughly the same number of onsets. In total, the recordings contain 1829 annotated onsets, while each instrument type contains roughly the same number of onsets. All recordings are monophonic, sampled at 44.1kHz. Requests for the dataset can be addressed to the second author.

B. Figures of Merit

For evaluating the results of the proposed onset detection system, the recall (R), precision (P), and F-measure (F) are employed as figures of merit. Let N_{tp}, N_{fp}, and N_{fn} stand for the number of correctly detected onsets, the number of false positives, and the number of missed onsets, respectively. Then, P and R are defined as:

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}}, \quad R = \frac{N_{tp}}{N_{tp} + N_{FN}} \tag{25}$$

while the F-measure is computed from P and R as:

$$F = \frac{2PR}{P + R} \tag{26}$$

It should be noted that P, R, and F are utilized for evaluation in the MIREX onset detection contests, as well as in [5], [7], [8]. An onset is correctly matched if it is detected within 50ms of the ground truth onset time, which is the same tolerance set for the MIREX onset detection task [28]. The presence of several detected onsets inside the 0.1s window counts as one correct detection, along with several false alarms. In the case of merged onsets, where an onset is detected within the tolerance window of two annotated onsets, one correct detection is reported along with a missed detection. By varying parameter δ in small steps, receiver operating characteristic (ROC) curves can be created by placing R values on the horizontal axis and P values on the vertical one [26]. The P/R-curve which is closer to the upper right corner of the diagram is considered to be the best detector with regards to F.

C. Results

The performance of the spectral flux, group delay, fundamental frequency estimator, and the fused system are shown in P/R curves in Fig. 7, for all instrument families. Regarding their optimum F-measure, the performance of the dataset is shown in Table II, for all onset strength signals. For the complete set of instruments the auditory spectral flux reports the highest result of 75.9%. However, the fusion of the three features yields a vastly improved rate of 82.6%. The achieved rate slightly outperforms the one in [5] by 0.5%, which also employed a fusion of the three descriptors without utilizing an auditory model. It should be noted that the fused onset strength signal for the complete set was created by combining the three descriptors using their best F-measure. However, results on the various instrument families for the fused descriptor were produced using the threshold values δ for the complete set. The performance of the isolated descriptors also outperforms the one reported in [5], where the auditory spectral flux has an improvement of 2% over the DFT-based spectral flux. Results using the auditory spectrum-based fundamental frequency estimator also report an improvement of 1.6% over the onsets derived by the YIN estimator.

Feature	SF	GRD	F_0	Fusion	COMP [1]	Fusion [5]
Bowed Strings	69.3%	67.6%	70.3%	77.6%	61.6%	75.4%
Pitched Percussive	84.4%	83.2%	75.4%	87.7%	86.3%	88.8%
Wind	73.3%	73.9%	79.9%	81.4%	71.7%	80.1%
All Instruments	75.9%	74.2%	75.7%	82.6%	73.2%	82.1%

TABLE II
BEST F-MEASURES FOR THE VARIOUS ONSET DETECTION FEATURES AND INSTRUMENT FAMILIES.

Feature	SF	GRD	F_0	Fusion
Western instruments	76.3%	77.0%	77.7%	83.6%
Middle-eastern instruments	76.1%	70.9%	72.6%	81.0%

TABLE III
BEST F-MEASURES FOR THE VARIOUS ONSET DETECTION FEATURES FOR WESTERN AND MIDDLE-EASTERN INSTRUMENT TYPES.

Concerning the statistical significance of the proposed method's performance compared to the method in [5], the recognizer comparison technique described in [29] was employed. The number of onset detection errors of the two methods is asumed to be distributed according to the binomial law. The average error rate of the method in [5] is $\hat{p}_1 = 0.1725$, while the average error rate of the proposed method is $\hat{p}_2 = 0.1465$. Taking into account that the test set size $s = 1829$ and considering 95% confidence ($\alpha = 0.05$), it can be seen that $\hat{p}_2 - \hat{p}_1 \geq z_\alpha \sqrt{2\hat{p}_1 s}$, where z_α can be determined from tables of the Normal law ($z_{0.05} = 1.65$). This indicates that the peformance of the proposed method is statistically significant compared to that in [5].

For comparative purposes, experiments were also performed using the method proposed by Bello et al. in [1]. The descriptor combining spectral and phase difference in the complex domain was computed using the MIR Toolbox [30]. Overall, the complex domain method reports an F-measure of 73.2% for the complete set, shown as COMP in Table II. Although the complex domain method reports results inferior to those reported using the auditory spectral flux and group delay for the complete set, the results using the pitched percussive set outperform all auditory spectrum-based and DFT-based approaches. The overall performance of the complex domain approach can be attributed to the fact that the onsets in the energy and phase-based functions are not aligned, since the phase-based approaches detect the actual start of the onset while energy-based approaches detect the largest energy increase in the attack part of the note. For pitched percussive instruments however, the attack part of the musical tone has shorter length compared to eg. string instruments, hence the two descriptors are aligned and reported rates tend to be high.

As far as the performance evaluation of the three instrument families is concerned, it can be seen that the best rate is achieved by the auditory spectral flux for the pitched percussive set, reaching 84.4%. However, the SF reports much lower rates for the bowed string and wind families. Comparing the performance of the DFT-based spectral flux and the auditory spectral flux for the pitched percussive set, it can be seen that the former slightly outperforms the latter, which can be attibuted to the lower spectral resolution of the auditory spectrum compared to the DFT. The auditory group delay (GRD) results are roughly the same compared to the spectral flux for each instrument family, however the onsets derived by the two descriptors are not identical, as is demonstrated by a fusion of the two aforementioned features: the best reported F-measure reaches 80.4% for the complete set, while the rates for the three instrument families are 74.5%, 88.4%, and 78.0%. It can be seen that the reported rate for the pitched percussive set outperforms the reported rate using the fusion of the three onset strength signals, signifying that the inclusion of F_0 estimation leads to inferior results for energy-based onsets, however greatly improves the detection of soft onsets. In general, the F_0 estimator outperforms energy and phase features for the string and wind families, reaching for the latter 79.9%. This is to be expected, since onsets created for string and wind instruments are often produced using constant excitation, with the only detectable change being in the pitch domain. Concerning the distinction between western and middle-eastern instruments (which are the bowed string kemençe, the wind ney, and the pitched percussive ud and tanbur), the reported F-measure for the western instruments using the fused descriptor is 83.6% and the respective for the middle-eastern instruments is 81.0%, as is shown in Table III. This is to be expected, since onsets for the ney and kemençe are occasionally not clearly defined and are more difficult to be estimated, even for a human annotator.

Addressing the performance depicted using the P/R curves in Fig. 7, it can be seen that the auditory group delay and F_0 estimator exhibit higher precision rates compared to the auditory spectral flux, which however compensates with higher recall values. For the pitched percussive set, the precision reached using the spectral flux and fused detectors reaches 100% for a recall rate of about 50%, making them ideal for beat tracking applications as in [9]. For the bowed strings set, a precision rate of about 90% is reported for the fused descriptor, for a recall of about 40%, which is also desirable since string onsets are the most difficult to detect. Also, using only the group delay feature, a precision of 87% is reached for a recall rate of 38%. Finally, using the F_0 estimator for the wind instruments set a precision of 87% is reported for a recall of 69%, making the F_0 estimator also suitable for beat tracking applications for wind instruments.

Some discussion on the performance of the various datasets reported in the literature should follow. For the pitched non-percussive and pitched percussive sets employed in [1], [2], [8], reported rates reach 98.4% in terms of F-measure for the pitched percussive data and 96.3% for the pitched non-percussive data. However, it should be noted that the dataset is less diverse, with the pitched percussive set consisting only of piano and guitar recordings and the pitched non-percussive set consisting of solo violin recordings, thus making a direct comparison of the rates between the two datasets impossible. Likewise, the pitch-based method proposed in [4] employs a dataset consisting of strings and singing voices with vibrato present, with a reported F-measure reaching 59.9% [31]. Thus, a fair comparison between onset detection methods can only be made using the same dataset.

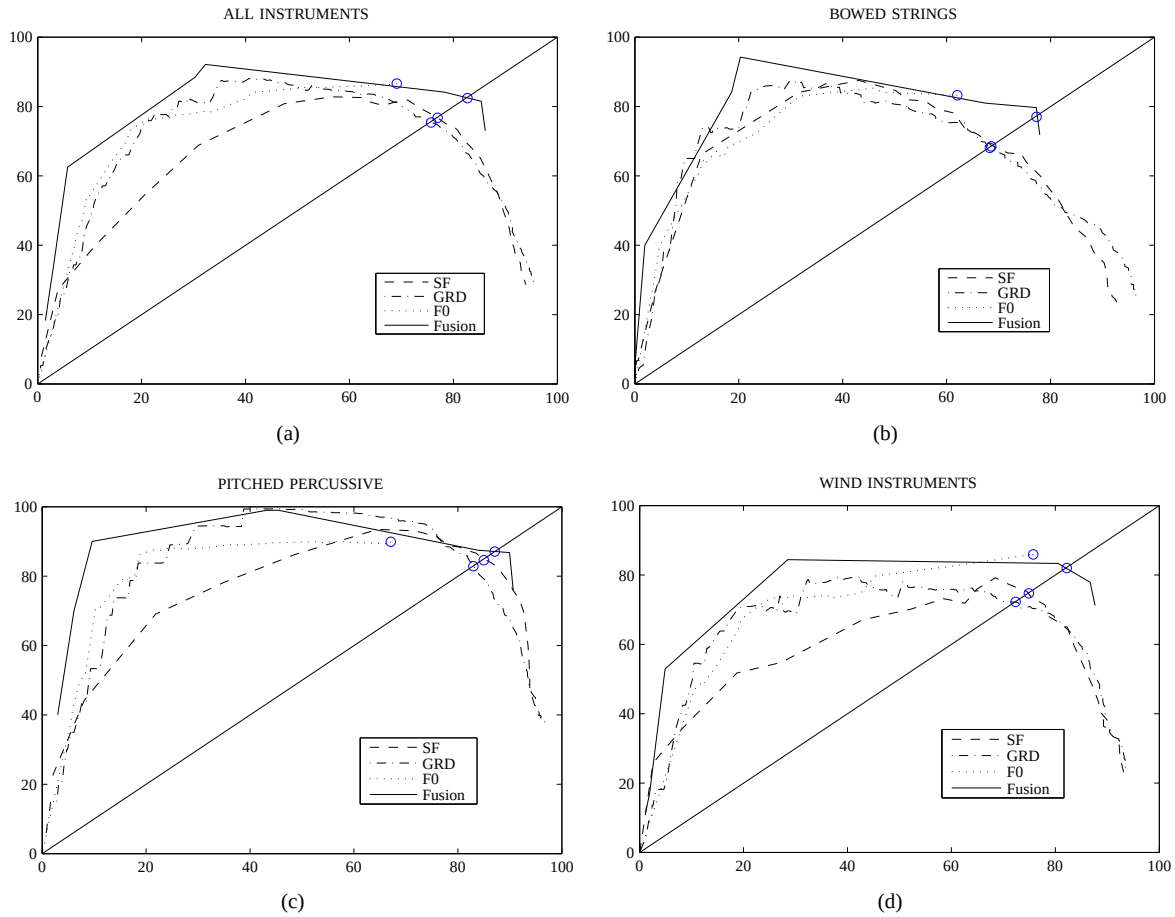

Fig. 7. Performance curves of the various onset detection descriptors. Recall and Precision values are plotted on the horizontal and vertical axis, respectively. A circle marker for each descriptor indicates the Recall-Precision pair which is closer to the upper left corner of the diagram.

V. Conclusions

In this paper, an approach for detecting onsets of pitched instrument recordings using auditory spectra was proposed. The group delay function and the spectral flux were derived for the auditory framework, and a novel fundamental frequency estimation algorithm using auditory spectra was presented. Experiments performed on a diverse dataset of pitched instrument recordings indicate that the auditory features show an improvement over standard state-of-the-art approaches for onset detection. In specific, the auditory spectral flux reached a high performance for detecting onsets of pitched percussive instruments, while a fusion of the spectral flux and group delay features at the decision level reached even better results. In addition, the auditory group delay and the fundamental frequency estimator report high precision rates for all instrument types, with the latter reaching high detection rates for string and wind instruments, whose onsets are more difficult to detect due to gradual energy changes. The combination of the three onset strength signals yields improved results, performing slightly better compared to the system proposed in [5], with the performance gain shown to be statistically significant.

In the future, the proposed algorithm for fundamental frequency estimation can be modified for multi-pitch estimation. In addition, the creation of an onset detection system which is dependent on the instrument family may possibly lead to improved results, as has been argued in previous MIREX competitions. The system could also consider onsets produced by non-pitched percussive instruments, which can be easily detected using energy descriptors. Finally, performance styles such as vibrato and ornamentations need to be taken into account for suppression in the creation of a truly robust onset detection system.

References

[1] 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 Proc. Letters*, Vol. 11, No. 6, pp. 553-556, June 2004.

[2] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, “A tutorial on onset detection of music signals,” *IEEE Trans. Speech and Audio Proc.*, Vol. 13, No. 5, pp. 1035-1047, Sep. 2005.

[3] D. Moelants and C. Rampazzo, “A computer system for the automatic detection of perceptual onsets in a musical signal,” in A. Camurri (Ed.), *KANSEI - The Technology of Emotion*, pp. 140-146, 1997.

[4] N. Collins, “Using a pitch detector for onset detection,” in Proc. *6th Int. Conf. Music Information Retrieval*, pp. 100-106, September 2005.

[5] A. Holzapfel, Y. Stylianou, A. C. Gedik, and B. Bozkurt “Three dimensions of pitched instrument onset detection,” *IEEE Trans. Audio, Language, and Speech Processing*, accepted for publication.

[6] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in Proc. *5th Int. Conf. Digital Audio Effects*, pp. 35-38, Sep. 2002.
[7] R. Zhou and J. D. Reiss, "Music onset detection combining energy-based and pitch-based approaches," in Proc. *8th Int. Conf. Music Information Retrieval*, Sep. 2007.
[8] S. Dixon, "Onset detection revisited," in Proc. *9th Int. Conf. Digital Audio Effects*, pp. 133-137, 2006.
[9] A. Holzapfel and Y. Stylianou, "Beat tracking using group delay based onset detection," in Proc. *9th Int. Conf. Music Information Retrieval*, Sep. 2008.
[10] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, Vol. 111, No. 4, pp. 1917-1930, 2002.
[11] R. Zhou and M. Mattavelli, "A new time-frequency representation for music signal analysis: resonator time-frequency image," in Proc. *9th Int. Symp. Signal Processing and Its Applications*, pp. 1-4, Feb. 2007.
[12] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in Proc. *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 3089-3092, March 1999.
[13] K. Jensen, "A causal rhythm grouping," in Proc. *2nd Int. Symp. Computer Music Modeling and Retrieval*, May 2004.
[14] N. Collins, "A comparison of sound onset detection algorithms with emphasis on phsychoacoustically motivated detection functions," in Proc. *AES 118th Convention*, May 2005.
[15] M. Gainza, E. Coyle, and B. Lawlor, "Onset detection using comb filters," in Proc. *IEEE Workshop Applications of Signal Processing to Audio and Acoustics*, pp. 263-266, 2005.
[16] P. Ru, "Multiscale multirate spectro-temporal auditory model," *PhD Thesis, Univ. Maryland College Park*, 2001.
[17] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Information Theory*, Vol. 38, No. 2, pp. 824-839, March 1992.
[18] T. F. Quatieri, *Discrete-Time Speech Signal Processing*, Prentice Hall, 2001.
[19] M. Brookes, P. A. Naylor, and J. Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Trans. Speech and Audio Proc.*, Vol. 14, No. 2, pp. 456-466, March 2006.
[20] A. V. Oppenheim, R. W Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, Prentice Hall, 1998.
[21] T. Chi and S. A. Shamma, "NSL Matlab Toolbox," http://www.isr.umd.edu/Labs/NSL/Software.htm, Neural Systems Lab., Univ. Maryland.
[22] A. Savitzky, and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry*, Vol. 36, No. 8, pp. 1627-1639, July 1964.
[23] I. Kauppinen, "Methods for detecting impulsive noise in speech and audio signals," in Proc. *14th Int. Conf. Digital Signal Proc.*, Vol. 2, pp. 967-970, July 2002.
[24] P. de la Cuadra, A. Master, and C. Sapp, "Efficient pitch detection techniques for interactive music," in Proc. *Computer Music Conf.*, Sep. 2001.
[25] A. M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum and a maximum likelihood estimate," in Proc. *Symposium Computer Processing in Communications*, Vol. XIX, Polytechnic Press: Brooklyn, New York, pp. 779-797, 1970.
[26] P. Leveau, L. Daudet, and G. Richard, "Methodology and tools for the evaluation of automatic onset detection algorithms in music," in Proc. *5th Int. Conf. Music Information Retrieval*, pp.72-75, 2004.
[27] K. Sjölander and J. Beskow, *Wavesurfer 1.8.5*, Centre for Speech Technology, KTH Royal Institute of Technology, `http://www.speech.kth.se/wavesurfer/`.
[28] MIREX 2009, "Audio onset detection task," http://www.music-ir.org/mirex/2009/index.php/Audio_Onset_detection .
[29] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, "What size test set gives good error estimates?," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, pp. 52-64, Jan. 1998.
[30] O. Lartillot and P. Toiviainen, "A MATLAB toolbox for musical feature extraction from audio," in Proc. *10th Int. Conf. Digital Audio Effects*, Sep. 2007.
[31] N. Collins, "Towards autonomous agents for live computer music: realtime machine listening and interactive music systems," *PhD Thesis*, Centre for Science and Music, Faculty of Music, University of Cambridge, 2006.

Emmanouil Benetos received the B.Sc. degree in Informatics in 2005 and the M.Sc. degree in Digital Media in 2007, both from the Aristotle University of Thessaloniki. In 2008 he was with the Multimedia Informatics Lab at the Department of Computer Science, University of Crete. Currently, he is pursuing his Ph.D. at the Department of Electronic Engineering and Computer Science, Queen Mary University of London, in the field of automatic music transcription. His research interests include music and speech processing, music information retrieval, and computational intelligence. He is a member of the Alexander S. Onassis Scholars' Association and a student member of IEEE.

Yannis Stylianou is Associate Professor at University of Crete, Department of Computer Science, CSD UOC and Associate Researcher in the Networks and Telecommunications Laboratory of the Institute of Computer Science at FORTH. He received the Diploma of Electrical Engineering from the National Technical University, N.T.U.A., of Athens 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. From 1996 until 2001 he was with AT&T Labs Research (Murray Hill and Florham Park, NJ, USA) as a Senior Technical Staff Member. In 2001 he joined Bell-Labs Lucent Technologies, in Murray Hill, NJ, USA (now Alcatel-Lucent). Since 2002 he is with the Computer Science Department at the University of Crete and the Institute of Computer Science at FORTH.

He 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. He is on the Editorial Board of Journal of Electrical and Computer Engineering, Hindawi (JECE), Associate Editor of the EURASIP Journal on Speech, Audio, and Music Processing, ASMP, and of the EURASIP Research Letters in Signal Processing, RLSP. He was Associate Editor for the IEEE Signal Processing Letters. He has over 100 peer-reviewed papers, and holds 9 US patents. He is member of IEEE, ISCA, and the Technical Chamber of Greece, TEE.