View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by TELKOMNIKA (Telecommunication Computing Electronics and Control)

Gamelan Music Onset Detection based on Spectral Features

Diah P. Wulandari, Aris Tjahyanto, Yoyon K. Suprapto
Institut Teknologi Sepuluh Nopember Surabaya
Kampus ITS Surabaya, 031-5947302 / 031-5931237
e-mail: diah@elect-eng.its.ac.id

Abstrak

Penelitian ini mendeteksi onset pada alat musik pukul dengan mengujinya pada sinyal suara alat musik gamelan sebagai salah satu alat musik tradisional Indonesia. Onset berperan penting dalam menentukan struktur ritmik dari sebuah musik, seperti beat, tempo, measure, dan sangat diperlukan dalam berbagai aplikasi untuk menemukan kembali informasi dalam sebuah potongan musik. Terdapat empat metode deteksi onset yang dibandingkan, yang memanfaatkan fitur-fitur spektral, yaitu magnitudo, fasa, dan kombinasi keduanya yaitu phase slope (PS), weighted phase deviation (WPD), spectral flux (SF), dan rectified complex domain (RCD). Fitur-fitur tersebut diekstrak dengan merepresentasikan sinyal suara ke dalam domain waktu-frekuensi menggunakan overlapped Short-time Fourier Transform (STFT) dengan menguji pengaruh panjang window. Fungsi deteksi onset diolah melalui sebuah proses pemilihan puncak menggunakan ambang batas dinamik. Hasilnya menunjukkan bahwa dengan pengaturan panjang window dan pengaturan parameter ambang batas dinamik yang sesuai dapat menghasilkan F-measure di atas 0,80 untuk beberapa metode tertentu.

Kata kunci: *deteksi onset, fitur spektral, musik gamelan, magnitudo, fasa*

Abstract

This research detects onsets of percussive instruments by examining the performance on the sound signals of gamelan instruments as one of traditional music instruments in Indonesia. Onset plays important role in determining musical rythmic structure, like beat, tempo, measure, and is highly required in many applications of music information retrieval. Four onset detection methods that employ spectral features, such as magnitude, phase, and the combination of both are compared in this paper. They are phase slope (PS), weighted phase deviation (WPD), spectral flux (SF), and rectified complex domain (RCD). Features are extracted by representing the sound signals into time-frequency domain using overlapped Short-time Fourier Transform (STFT) and by varying the window length. Onset detection functions are processed through peak-picking using dynamic threshold. The results showed that by using suitable window length and parameter setting of dynamic threshold, F-measure which is greater than 0.80 can be obtained for certain methods.

Keywords*: onset detection, spectral features, gamelan music, magnitude, phase.*

1. Introduction

Onset in music is considered as the abrupt change which marks the start of a note transient [1].Clapping hands, tapping feet, or even moving bodies while listening to music is caused by the ability of human being to perceive onsets in music [2]. Detection of onsets influences the recognition of other musical features such as beat, tempo, and rhythm. It is even useful for many high level applications in music information retrieval, like instrument identification, and musical fingerprinting. Various methods have been proposed to detect onsets in music signals. Most of the methods have been tested on western music containing the sounds of Musical Instruments Digital Interface (MIDI) and fabricated acoustic instruments, like organ, piano, guitar, and violin but only few of them were tested on traditional music [2]. Traditional music has their own characteristics that discriminate them from the other types of music. Therefore, onset detection in traditional music signals has specific challenge due to the uniqueness of the instruments, rules and playing styles.

Gamelan is one of Indonesia's traditional music instruments originated from Jawa and Bali. One gamelan set consists of about fourteen different types of instruments [3]. Most of the

Received September 3, 2012; Revised January 3, 2013; Accepted January 28, 2013

instruments are percussive. There are many variations in gamelan music signals that are induced by three factors. First, the construction process of gamelan involves unstandardized tools and materials. The constructors usually tune a new instrument by comparing its sound with that of the tuner using their ears. Second, there are different rules for gamelan in different regions like Semarang, Yogyakarta, Solo, and Jawa Timur. The fundamental frequency of instrument notations may be slightly different among those regions. Third, gamelan has a unique playing rule compared to that of western music. Unlike the pattern of western music which flows like water from upstream going down the river, the playing pattern of gamelan music is repetitive in the sense that the musicians may repeat the pattern as long as needed, for example as accompaniment of traditional wedding ceremony and art show. Western music also employs fix and uniform rules and standards for its instruments, including the tools and materials used in the construction process. That is why western music signals have less variations compared to those of traditional music signals, specifically gamelan music signals. This research focuses on the analysis of gamelan music onset detection containing the sounds of *demung*, *saron*, *peking*, representing balungan family and *bonang* from gong family, which are the common instruments found in an ensemble.

Onset detection methods mainly comprise three steps, which are feature extraction, reduction function, and peak-picking, as depicted in Figure 1 [1-4]. First, features of audio signal are extracted based on the analysis domain. Next, the features are fed into reduction function which generates onset detection function. Finally, a threshold function is applied on the onset detection function in order to select onsets among the candidates. Reduction function is the step that differentiate an approach from the others. There are three major approaches in the way people develop the reduction function of onset detection method. The first two major approaches are those based on signal features and those based on probability model [1]. The last one is data driven reduction function [4].

People exploited the features of audio signals both in time domain and spectral domain (time and frequency domain). Since the occurrence of onset is usually accompanied by an increase of amplitude, people developed an envelope follower by rectifying and smoothing the signal [5]. This is mostly suitable for signal with strong percussive sounds. A refinement of this method used the first time-derivative of energy which is usually combined either with filter-banks or transient-steady state separation [6-8].

Another idea that people have been developed is to represent signal in time-frequency domain and to utilize the spectral structure to build onset detection function. A number of transformation tools have been used to produce such structure, like Short-time Fourier Transform (STFT), wavelet transform, constant-Q transform, etc. This method is suitable for signals with multiple instruments. Seeing onsets as points in an N-dimensional space, some researches produced onset detection function based on *spectral difference*. They measured the distance between successive spectra using general distance metrics, such as L_1-norm [9] and L_2-norm [10]. A method *called high frequency content* (HFC) was also proposed by Masri [9]. It applied linear weights to each frequency bin of the spectral structure. This approach aims to balance the energy across the frequency by emphasizing it in high frequency, where it is likely to be less concentrated.

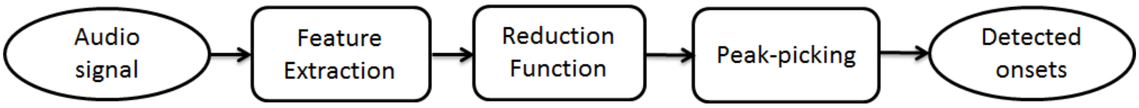

Figure 1. Onset detection method in general

A report made by Davies et. al. [11] on the decreasing performance of state-of-the-art approach of beat tracking when dealing with signals with weak percussive contents. Therefore they exploit the phase spectrum of signals. The first derivative of phase over frequency, namely *group delay*, was used in [12]. The onset detection function was generated by taking the average of group delay function over frequency resulting *phase slope* function. Onsets were estimated by detecting the positive zero crossings of this function. This method does not require threshold function, but it only worked well for synthetic signals. It may encounter problem with real signals where noise is present. Another method which employs the phase spectrum of signals calculates the *instantaneous frequency* by taking the first time-derivative of phase [13] [1]. The change of instantaneous frequency indicates possible onsets. The onset detection function is defined as the mean of absolute change in instantaneous frequency. Improvements of this method are *weighted phase deviation*, where each frequency bin was weighted by its magnitude.

Both amplitude and phase can be predicted based on their states on the previous two bins, as presented in [14], [1]. The absolute deviations between the expected values and the real ones produce the onset detection function of *complex domain* method. A research has been conducted to investigate the performance of onset detection method by exploiting the spectral features [12]. Each method was tested on dataset with *pitched non-percussive* instruments, *pitched percussive* instruments, *non-pitched percussive* instruments, and *complex mixture*. The result showed that *spectral flux* and *complex domain* methods were best for *pitched percussive* instruments, while *phase deviation* method was best for *non-pitched percussive* ones. All possible combinations of three spectral features, which are *phase slope*, *spectral flux*, and *fundamental frequency change* have been tested to detect onsets on data set containing *percussive* instruments, *bowed* instruments, *wind* instruments, and *complex mixture* [2]. Experiments proved that the fusion of all features produced significant increase of F-measure, compared to other combinations.

The other two main approaches in onset detection task are those which employ data driven reduction function and probabilistic reduction function. As the name implies, data driven method requires a large number of data to train the discriminate function to recognize onsets. It is based on statistical model available in dataset, and generates a model through learning process. The most common methods are using Gaussian mixture model [15], support vector machine [16] and neural networks [3], [17-19].The last approach constructs probability inference about the likely times of onsets based on some observations [1]. People also used divergence algorithm to detect onsets based on sequential probability ratio test [20]. While other research build onset detection function by calculating the negative log probability of the signal given its recent history [21]. Probabilistic approach provides more general theoretical concept of onsets [22], but the model built may suffer from the complexity of the signal due to the diverse kinds of instrument played. Therefore this approach is most suitable for detecting onset of a single instrument.

Due to the special characteristics and diversity of gamelan music signals, we investigate the performance of onset detection methods exploiting signal's spectral features. Methods based on magnitude, phase and their combination that have been proposed previously, are tested on gamelan music signals. Dataset consists of the recordings of real gamelan playing. The content of this paper is organized as follows. Section 1 explains some backgrounds of gamelan music onset detection along with related researches. Section 2 describes the methods used in this research and the experimental settings. Section 3 presents and analyzes about the results of experiments. And finally, section 4 draws conclusions based on the analysis.

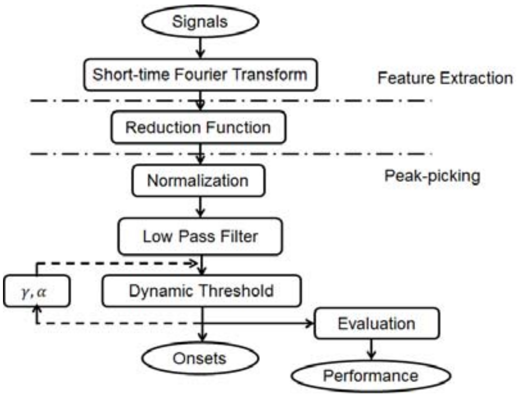

Figure 2. Block diagram of the experiment

2. Research Method

This section describes fundamental theories for the methods proposed in this paper including equations and all supporting materials. The explanation flows like the block diagram of onset detection method along with pre-processing and post-processing steps carried out in this research, as depicted in Figure 2. It starts from STFT for feature extraction. Next, it continues with one-by-one experiment of four reduction functions exploiting the magnitude, phase, and combination of both features. Finally, it describes post-processing steps that comprise normalization, low pass filtering and dynamic thresholding. Two parameters were adjusted in dynamic threshold in order to obtain the best performance. Details of each block are described in the following sub sections.

2.1. Short-time Fourier Transform (STFT)

Discrete Fourier Transform (DFT) provides signal analysis in frequency domain. It reveals the frequency content within the signals. In this way, Fourier theory assumes that the signal is stationary in terms of mean, power, power spectrum, and other statistic components [21], while most of signals in nature like speech, music, and image are non-stationary ones. Therefore a short-time analysis of the signals is required. STFT is constructed by computing successive DFT frames for input signals [23].

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

n and ω are time index and frequency index respectively. The frames are obtained by applying common window functions, such as Hamming, Hann, or rectangular windows. The chosen window length affects both time and frequency resolutions. According to Heisenberg uncertainty principle [24] [25], time resolution of STFT is reversely proportionate to its frequency resolution, and is shown in eq. (2).

$$\Delta t\Delta f\leq\frac{1}{4\pi} \tag{2}$$

Δt and Δf are time resolution and frequency resolution respectively and the minimum value of $\Delta t\Delta f$ is known as *Heisenberg box*. The use of longer window increases frequency resolution but at the same time it decreases time resolution. One way to improve time resolution while maintaining frequency resolution is by applying overlapped STFT and it can be calculated using eq. (3).

$$\Delta t=\frac{l_h}{f_s} \tag{3}$$

l_h is the hop length of overlapped STFT and f_s is sampling frequency.

2.2. Reduction Functions

This sub section exposes the concepts behind each method. The first two methods are based on phase spectrums, the third is based on energy, and the last is based on combination of magnitude and phase.

2.2.1. Phase Slope (PS)

The first frequency-derivative of phase spectrum is considered as *group delay*. The Fourier transform of a delayed unit sample sequence $x[n]=\delta[n-n_0]$ is $X(\omega)=e^{-j\omega n_0}$ and according to eq. (3) the group delay of such signal is $\tau(\omega)=\omega n_0\ \forall\omega$ [23].

$$\tau(\omega,n)=-\frac{d\varphi(\omega,n)}{d\omega} \tag{4}$$

Equation (4) shows the calculation of group delay function from the time-frequency representation of a signal. $\varphi(\omega,n)$ is the unwrapped phase spectrum of STFT. Taking the average of group delay over frequency gives the negative of the slope of the phase spectrum and is equal to n_0 in the case of delayed unit sample sequence, as mentioned in eq. (5).

$$PS(\omega,n) = \frac{1}{N}\sum_{\omega=-\frac{N}{2}}^{\frac{N}{2}-1}\tau(\omega,n) \tag{5}$$

N is the window length. The phase slope function represents the distance between the center of the analysis window and the position of the impulse. This leads to a conclusion that a zero-crossing of phase slope function indicates zero distance to the impulse and it can be used to detect onsets.

2.2.2. Weighted Phase Deviation (WPD)

It is an improvement of a method that based the onset detection on the rate of change of phase in STFT frequency bin, called *phase deviation*. It is an estimate of the instantaneous frequency of that component [12]. The weighting was meant to enhance the peaks with high energy and reduce the susceptibility of the method to noise. First the instantaneous frequency is calculated by taking the first time-derivative of the phase spectrum of STFT. The change of instantaneous frequency is an indication of possible onsets. Then the second time-derivative of phase spectrum is multiplied by its magnitude, as presented by eq. (6).

$$WPD(\omega,n) = \frac{1}{N}\sum_{\omega=-\frac{N}{2}}^{\frac{N}{2}-1}|X(\omega,n)\varphi''(\omega,n)| \tag{6}$$

$\varphi(\omega,n)$ is the second time derivative of unwrapped phase spectrum.

2.2.3. Spectral Flux (SF)

This feature represents the change of magnitude over time for each frequency bin. The onset detection function is defined as the positive change of spectral flux summed across all frequency bins, as described by eq. (7) and eq. (8).

$$SF(n) = \sum_{\omega=-\frac{N}{2}}^{\frac{N}{2}-1}H(|X(\omega,n)| - |X(\omega,n-1)|) \tag{7}$$

$$H(x) = \frac{x+|x|}{2} \tag{8}$$

where $H(x)$ is the half wave rectifier function. This method highly bases the detection on the energy of signals and therefore is suitable for signals containing percussive contents.

2.2.4. Rectified Complex Domain (RCD)

Complex domain method considers both spectrums of signals which are magnitude and phase. It generates an expectation of magnitude and phase based on the values of those two component in previous two bins [13]. This target is produced by assuming constant magnitude and constant rate of phase change, according to eq. (9).

$$X_T(\omega,n) = |X(\omega,n-1)|e^{\varphi(\omega,n-1)+\varphi\prime(\omega,n-1)} \tag{9}$$

In order to distinguish between increases and decreases of signal amplitude, the concept of half wave rectifier is used, resulting in the function $RCD(\omega,n)$. The rectification only considers deviation between real and target values when $|X(\omega,n)| \geq |X(\omega,n-1)|$ as mentioned in eq. (10). The $RCD(\omega,n)$ function is then summed up across frequency to build the onset detection function.

$$RCD(\omega,n) = \begin{cases} |X(\omega,n) - X_T(\omega,n)| & if\ |X(\omega,n)| \geq |X(\omega,n-1)| \\ 0 & otherwise \end{cases} \tag{10}$$

$$RCD(n) = \sum_{\omega=-\frac{N}{2}}^{\frac{N}{2}-1}RCD(\omega,n) \tag{11}$$

2.3. Normalization

The signal’s energy among different instruments may vary one another. For percussive instruments, the energy highly depends on the strength applied while hitting the instruments. Normalization is one way to encounter such condition and to help the thresholding process. The onset detection function is normalized by subtracting the mean and dividing the result by the maximum value of the function.

$$\bar{d}(n) = d(n) - \frac{1}{N}\sum_{i=1}^{N} d(n_i) \tag{12}$$

$$\tilde{d}(n) = \frac{\bar{d}(n)}{max\{|\bar{d}(n)|\}} \tag{13}$$

Equation (12) and equation (13) describe the normalization of onset detection function $d(n)$, where$max\{.\}$ is the maximum function.

2.4. Low Pass Filter

This is the process carried out to suppress noise and high frequency ripples resulted from feature extraction. Kaiser window was chosen since it has most energy at main lobe for a given side lobe amplitude. The design of the filter was adjusted to three parameters, which are sampling frequency of onset detection function F_{so}, passband frequency F_{pass}, and stopband frequency F_{stop}. F_{so} depends on the window length and hoplength of STFT, while F_{pass} is the approximate frequency of onset occurences, and F_{stop} is the approximate frequency of ripples.

$$F_{so} = \begin{cases} \frac{F_s}{l_w} & for\ non - overlap\ STFT \\ \frac{F_s}{l_h} & for\ overlap\ STFT \end{cases} \tag{14}$$

Equation (14) shows how to derive the sampling frequency of onset detection function, where l_w and l_h are window length and hop length of STFT respectively.

2.5. Dynamic Threshold

Static or fixed thresholding is the simplest way to help performing peak-picking. But it is not a wise choice since setting higher threshold value may increase the number of false negative (missed detection) while setting lower threshold value may increase the number of false positive. It is highly recommended to use dynamic thresholding, especially for signals with high variations.

Unlike the static one, dynamic thresholding may adjust the threshold value based on signal statistics. There are several methods of thresholding as mentioned in [1] and [13]. The method used in this paper exploits the median filter, like the one stated in [1]. Details of dynamic threshold function are described by eq. (15).

$$\tilde{\delta}(n) = \gamma + \alpha \times median\{|d(n-M)|, \ldots, |d(n+M)|\} \tag{15}$$

$\tilde{\delta}(n)$is dynamic thresholding function, M is half the window of median filter$median\{.\}$, whileγ and α are constants adjusted to get the best result. Median filter is a function which outputs the median value of the samples inside a certain window. The value of $M$is following eq. (16).

$$M = \frac{1}{2} \times \frac{F_{so}}{F_{onset}} \tag{16}$$

A peak is selected as onset among the candidates if it satisfies the conditions mentioned in eq. (17) - (19):

$$\tilde{d}(n) < \tilde{d}(n-1) \tag{17}$$

$$\tilde{d}(n) < \tilde{d}(n+1) \tag{18}$$

$$\tilde{d}(n) > \tilde{\delta}(n) \quad (19)$$

2.6. Performance Evaluation

Performance of onset detection task is evaluated using $F-measure$, a common measure for this task based on Music Information Retrieval and Exchange (MIREX) [2] [4] [12]. Three parameters are measured, which are true positive (TP), false negative (FN), and false positive (FP). TP represents the number of correctly detected onsets, FN represents the number of undetected onsets, and FP represents the number of non-onset events which are detected as onsets. These parameters are used to calculate precision (P), and recall (R). Equation (20)-(22) details the calculation of performance measure.

$$P = \frac{TP}{TP+FP} \quad (20)$$

$$R = \frac{TP}{TP+FN} \quad (21)$$

$$F = \frac{2\times P\times R}{P+R} \quad (22)$$

3. Results and Discussion

The objective of the experiments is to compare the performance of each of the four methods on specific composition of instrument in gamelan music. Data were taken from original recordings of the playing of one gamelan set. Each method was tested on gamelan music signals containing single instrument playing and also on the mix of two instruments. Evaluation was carried out on the results using $F-measure$. In the first sub section, the composition of dataset is explained as well as the procedure and parameter setting of feature extraction. The next subsection presents the comparison of performance of each method. The analysis also investigates the effect of STFT window length on the performance of each method.

3.1. Data Set

Data set can be divided into two compositions of instruments, which are single instrument and mixed instruments. Single instrument composition consists of bonang, demung, saron, and peking; while mixed instrument composition consists of saron-bonang, saron-demung, and saron-peking. The aim of such division is to examine and compare the ability of each onset detection method on different All data are originally recorded from real playing of one gamelan orchestra. Each of the raw signals is extracted using overlapped STFT, in order to provide 10 ms time resolution on 48 kHz sampling frequency.

Figure 3 depicts gamelan signals of an excerpt of Javanese song called Manyar Sewu. Each graph represents signal containing the sounds of one gamelan instrument, which are bonang, demung, peking and saron.

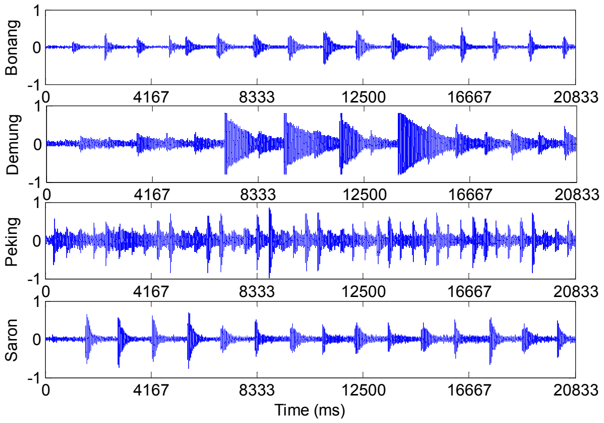

Figure 3. Gamelan instrument sound signals

We notice that bonang produces signals with relatively smaller amplitude and shorter duration, while peking produces short duration but with bigger amplitude than that of bonang. The amplitude varies among different type of instruments, and it is also fluctuating among different notations (sound source) in one instrument, as clearly shown by the signal of demung. This represents variations in gamelan music signals which provides challenge on onset detection task.

3.2. Comparison of All Methods

Figure 4 and Figure 5 show onset detection function resulted from phase slope (PS), weighted phase delay (WPD), spectral flux (SF), and rectified complex domain (RCD) on an excerpt of gamelan music signal. The effect of window length was investigated in the experiments, using 2048 and 8192 samples. Table 1 and Table 2 numerically summarize the performance measures for all methods using different window length of STFT. From our observation on gamelan music signals, we assume that in average there are three onsets in one second F_{onset} = 3 Hz, therefore we set F_{pass} = 2 Hz, and F_{stop}= 5 Hz for the low pass filter.The window length of median filter was set to be ±80 ms according to the calculation in eq. (16).

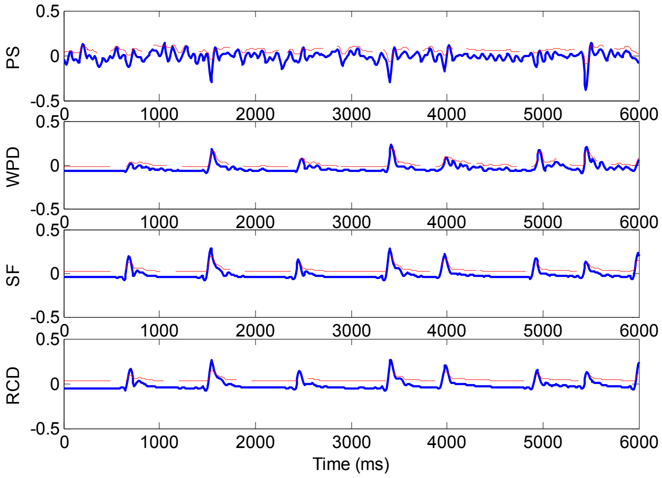

Figure 4. Onset detection functions (solid lines) obtained from each of four methods as well as threshold functions (dashed lines) using 2048 window length and 10 ms overlapped STFT

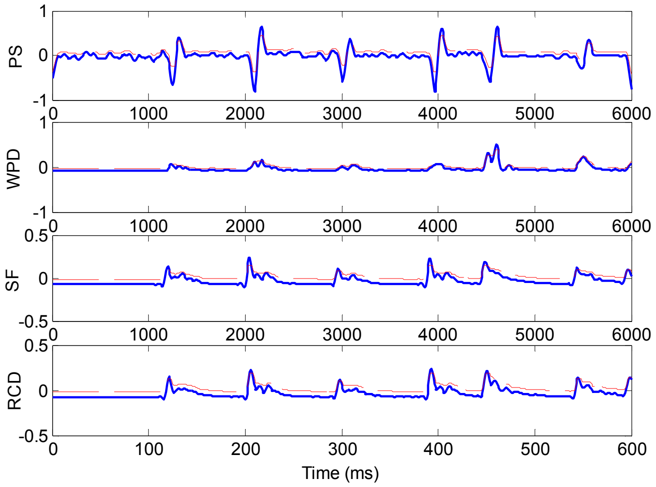

Figure 5.Onset detection functions (solid and bold lines)obtained from each of four methods as well as threshold functions (dashed lines)using 8192 window length and 10 ms overlapped STFT

Onsets were annotated by considering ± 25 ms tolerance window. This means that if there are more than one onset occurred inside this length of window, only one of them is assumed to be the true onset. Such condition usually happens on signals containing two instruments, for example saron and demung. Since these two instruments have the same notations, they tend to be hit at the same time. This causes onset events with short time interval (<100 ms). We observed that multiple onsets occurred with time interval less than 50 ms cannot be distinguished by human eyes nor by human ears. This tolerance window was also applied during performance evaluation by taking into account delay time caused by low pass filter. The parameters γ and α were adjusted independently for each method in order to obtain the optimal performance.

From Fig. 4 and Fig. 5, we notice that PS showed improvement using longer window while the rest three methods yielded better result using shorter window. In the case of PS, since onset events are marked based on zero-crossing, this method is sensitive to noise which cannot be avoided in real recording data. Therefore the onset detection function of PS showed many ripples which came from the noise presented in the signals. The use of longer window was able to suppress this noise and to enhance the onset strength signals at the same time. The performance of PS by using 8192 window length reached twice of that by using 2048 window length, as shown in Table 1 and Table 2. Threshold function was employed to reduce over detection on noise.

Table 1. Performance evaluation of onset detection task using 2048 window length STFT

No	Instrument composition	No. Of True Onsets	Method	Precision	Recall	F-measure
1.	Bonang	144	PS	0,44	0,43	0,43
			WPD	0,66	0,67	0,66
			SF	**0,86**	**0,86**	**0,86**
			RCD	0,84	0,83	0,83
2.	Demung	44	PS	0,56	0,57	0,56
			WPD	0,43	0,41	0,42
			SF	**0,73**	**0,73**	**0,73**
			RCD	0,64	0,66	0,65
3.	Saron	74	PS	0,65	0,66	0,66
			WPD	0,70	0,70	0,70
			SF	**0,92**	**0,92**	**0,92**
			RCD	0,89	0,91	0,90
4.	Peking	98	PS	0,74	0,73	0,74
			WPD	0,88	0,88	0,88
			SF	**0,99**	**0,99**	**0,99**
			RCD	0,97	0,96	0,96
5.	Saron-bonang	66	PS	0,52	0,53	0,53
			WPD	0,78	0,80	0,79
			SF	0,99	1,00	0,99
			RCD	**1,00**	**1,00**	**1,00**
6.	Saron-demung	65	PS	0,64	0,63	0,64
			WPD	0,79	0,77	0,78
			SF	**0,97**	**0,97**	**0,97**
			RCD	**0,97**	**0,97**	**0,97**
7.	Saron-peking	150	PS	0,70	0,71	0,70
			WPD	0,85	0,85	0,85
			SF	**0,93**	**0,92**	**0,93**
			RCD	0,93	0,92	0,92

Table 2. Performance evaluation of onset detection task using 8192 window length STFT

No	Instrument composition	No. Of True Onsets	Method	Precision	Recall	F-measure
1.	Bonang	144	**PS**	**0,86**	**0,85**	**0,86**
			WPD	0,62	0,62	0,62
			SF	0,76	0,75	0,76
			RCD	0,74	0,75	0,75
2.	Demung	44	**PS**	**0,91**	**0,93**	**0,92**
			WPD	0,43	0,43	0,43
			SF	0,57	0,57	0,57
			RCD	0,64	0,66	0,65
3.	Saron	74	**PS**	**0,92**	**0,94**	**0,93**
			WPD	0,68	0,69	0,68
			SF	0,84	0,86	0,85
			RCD	0,84	0,86	0,85
4.	Peking	98	**PS**	**0,99**	**0,99**	**0,99**
			WPD	0,71	0,71	0,71
			SF	0,89	0,89	0,89
			RCD	0,86	0,86	0,86
5.	Saron-bonang	66	PS	0,90	0,92	0,91
			WPD	0,63	0,62	0,63
			SF	**0,96**	**0,97**	**0,96**
			RCD	0,82	0,83	0,83
6.	Saron-demung	65	PS	0,92	0,92	0,92
			WPD	0,71	0,74	0,72
			SF	**0,97**	**0,97**	**0,97**
			RCD	**0,97**	**0,97**	**0,97**
7.	Saron-peking	150	**PS**	**0,89**	**0,89**	**0,89**
			WPD	0,66	0,67	0,67
			SF	0,74	0,73	0,74
			RCD	0,66	0,66	0,66

The other three methods (WPD, SF, and RCD) obtained better results using shorter window since the onset detection functions produced by these methods took the derivative of the spectrum over time. STFT using longer window tends to make the spectrum smoother since it considers longer portion of signal inside the window. By using the same hop length, window length to hop length ratio $(l_w\colon l_h)$ is greater for STFT with longer window length, thus the spectrums resulted from the slided windows have small differences each another which causes smoother spectrum surface. At one side, this effectively reduces noise but on the other side this causes low time derivative of the spectrum and therefore signal transients are hard to detect. PS is the only method which builds its onset detection function by taking the derivative of its spectrum over frequency, therefore the use of longer STFT window successfully suppressed the noise in the signal without losing information of onset peaks.

Table 3. Overall performance of the methods

2048		8192	
Method	**F-measure**	**Method**	**F-measure**
PS	0,61	PS	0,91
WPD	0,75	WPD	0,65
SF	0,92	SF	0,80
RCD	0,90	RCD	0,78

By comparing precision and recall values of all experiments presented in Table 1 and Table 2, we may conclude that for all methods, precision and recall ratio is approximately 1 $(\mathrm{P}\colon\mathrm{R} \approx 1)$, showing that they are of the optimal conditions [2]. The parameter setting of dynamic threshold played the role to balance precision and recall of onset detection. The winning method for each composition is shown by the bold values in shaded rows. In Table 1, SF reached the best F-measure for all instrument compositions except saron-bonang. While in Table 2, PS won the best F-measure for bonang, demung, saron, peking, and saron-peking. PS left the other methods with significant difference of F-measure. The performance of RCD followed that of SF in Table 1 and beated the other methods for saron-bonang and saron-demung compositions, but it only got the best result for saron-demung composition in Table 2, all with a slight difference of performance. WPD outperformed PS in Table 1, but it was on the lowest place in Table 2.

Although theoretically PS is stated as a method which is able to detect soft onsets, it failed to perform well in bonang signals. This was caused by the presence of noise and the sound of other instrument in the signal. PS reached the best performance on single peking since the signals of peking have relatively high amplitude and short duration so the occurrence of onsets can be distinguished clearly. WPD obtained the best performance on single peking using short window but it reached the best performance on saron-demung composition using long window as well as the other two methods (SF and RCD).Signals of saron and demung have higher amplitudes than the other instruments (bonang and peking) and data containing the composition of saron and demung have stable amplitude compared to data containing single saron and single demung as shown in Fig. 3. Therefore SF and RCD obtained the lowest performance dealing with single demung composition using both window lengths.

The overall performance of all methods on all dataset is presented in Table 3. From these results, we conclude that SF is robust to the STFT window length and the instrument composition in the signal. It obtained F-measure greater than 0,80 using both window lengths. RCD follows SF in with small difference of F-measure (≈ 0,02). PS is more stable using longer STFT window regardless the instrument compositions with F-measure greater than 0,85. It outperformed the other methods with significant difference of F-measure (≈ 0,1). The use of dynamic threshold also contributed to the robustness of the system.

4. Conclusions

From the results we may derive three conclusions. First, PS produces better performance using long window while the others (WPD, SF, and RCD) works better using short window. This condition is influenced by the way each method develops the onset detection function and by the ratio of window length and hop length used to extract features. Methods which build onset detection function by taking time derivative of the spectrum give higher performance using shorter STFT window. While methods which build onset detection function by taking frequency derivative give lower performance using longer STFT window. The experiment results also support the conclusion in [2] that PS method is most suitable for pitched percussive instruments, like those in gamelan. Second, variations contained in gamelan music signals and natural condition occurs in original recordings of gamelan music (i.e. noise) affects the performance of each method on onset detection task. For example, the variations of amplitude and spectral envelope contributed to the number of false negative and false positive. Third, the use of dynamic threshold with suitable parameter setting help the onset detection methods encountering such variations in gamelan music signals thus supports the performance of all methods.

These methods should be tested on more complex gamelan music signals, those which contain the playing of gamelan orchestra (more than two instruments). The difficulty of handling these kinds of audio music signals is on how to annotate the onsets. Several gamelan musicians must be involved to conduct hand labeling of onsets. The onset detection function will be used in many other applications, like beat tracking, measure estimation, and music transcription.

References

[1] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, M. B. Sandler. A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing.* 2005; 13(5): 1035-1047.

[2] A. Holzapfel, Y. Stylianou, A. C. Gedik, B. Bozkurt. Three Dimensions of Pitched Instrument Onset Detection. *IEEE Transactions on Audio, Speech, and Language Processing*. 2010; 18(6): 1517-1527.

[3] Sumarsam. Cultural Interaction and Musical Development in Central Java. Chicago, Illionis, USA: The University of Chicago Press. 2003. 333-342.

[4] F. Eyben, S. Bock, B. Schuller, A. Graves. *Universal Onset Detection with Bidirectional Long Short-term Memory Neural Networks*. 11th International Society for Music Information Retrieval. Utrecht, Netherlands. 2010. 589-594.

[5] A. W. Schloss. On the Automatic Transcription of Percussive Music – From Acoustic Signal to High-Level Analysis. Ph.D. Dissertation. Tech. Rep. STAN-M-27. Dept. Hearing and Speech, Univ. Stanford, CA. 1985.

[6] M. Goto, Y. Muraoka. *Beat Tracking based on Multiple-agent Architecture – a Real-time Beat Tracking System for Audio Signals*. 2nd International Conference of Multiagent Systems. December 1996. 103-110.

[7] S. Levine. Audio Representations for Data Compression and Compressed Domain Processing. Ph.D. dissertation. Univ. Stanford, CA. 1998.

[8] C. Duxbury, M. Davies, M. Sandler. *Improved time-scaling of musical audio using phase locking at transients*. 112th AES Conference. Munich, Germany. 2002. 5530.

[9] P. Masri. Computer Modeling of Sound for Transformation and Synthesis of Musical Signal. Ph.D. Dissertation. Univ. Bristol, U.K. 1996.

[10] C. Duxbury, M. Sandler, M. Davies. *A Hybrid Approach to Musical Note Onset Detection*. Digital Audio Effects Conference (DAFX'02). Hamburg, Germany. 2002: 33-38.

[11] M. E. P. Davies, M. D. Plumbley. Context-dependent beat-tracking of musical audio. *IEEE Transactions on Audio, Speech, and Language Processing*. 2007. 15(3): 1009-1020.

[12] S. Dixon. *Onset Detection Revisited*. 9th International Conference on Digital Audio Effects. Montreal, Canada. 2006. 133-137.

[13] C. Duxbury, J. Bello, M. Davies, M. Sandler. *A Combined Phase and Amplitude based Approach to Onset Detection for Audio Segmentation*. 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-03). 2003. 275-280.

[14] J. Bello, C. Duxbury, M. Davies, M. Sandler. On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain. *IEEE Signal Processing Letters.* 2004; 11(6): 553-556.

[15] Y. Wei, R. B. Dannenberg. *Polyphonic Music Note Onset Detection using Semi Supervised Learning*. Austrian Computer Society. 2007.

[16] C. Toh, B. Zhang, Y. Way. *Mulitple Feature-fusion based Onset Detection*. International Seminar on Music Information Retrieval. 2008.

[17] A. Lacoste, D. Eck. Onset Detection with Artificial Neural Networks. Music Information Retrieval and Exchange. 2005.

[18] Y.Tanoto, W.Ongsakul, C.O.P. Marpaung. Levenberg-Marquardt Recurrent Networks for Long-term Electricity Peak Load Forecasting. *Telkomnika Journal*. 2011; 9(2).

[19] N. R.Emillia, NFN Suyanto, W. Maharani. Isolated Word Recognition using Ergodic Hidden Markov Models and Genetic Algorithm. *Telkomnika Journal*. 2012; 10(1).

[20] M. Basseville, A. Benveniste. Sequential Detection of Abrupt Changes in Spectral Changes of Digital Signals. *IEEE Transactions on Information Theory*. 1983: 29.

[21] S. A. Abdallah, M. D. Plumbley. *Probability as Metadata: Event Detection in Music using ICA as a Conditional Density Model*. 4th International Symposium on Independent Component Analysis and Signal Separation (ICA 2003). Nara, Japan. 2003: 233-238.

[22] S. V. Vaseghi. Multimedia Signal Processing: Theory and Applications in Speech, Music, and Communications. West Sussex, UK: John Wiley & Sons, Ltd. 2007: 57-59.

[23] T. H. Park. Introduction to Digital Signal Processing: Computer Musically Speaking. Singapore. World Scientific, Co., Pte., Ltd. 2010: 292-293.

[24] S. G. Mallat. A Wavelet Tour of Signal Processing. New York: Academic. 1998.

[25] J. Zhong, Y. Huang. Time-Frequency Representation Based on an Adaptive Short-time Fourier Transform. *IEEE Transactions on Signal Processing.* 2012; 58(10).

[26] A. Holzapfel, Y. Stylianou. *Beat Tracking using Group Delay based Onset Detection*. 9th International Seminar on Music Information Retrieval. Pennsylvania, USA. 2008. 653-658.