❐   1700

Impulsive spike enhancement on gamelan audio using harmonic percussive separation

Solekhan[1], Yoyon K. Suprapto[2], Wirawan[3]
[1,2,3]Department of Electrical Engineering, Faculty of Electrical Technology,
Institut Teknologi Sepuluh Nopember, Indonesia
[1]Department of Electrical Engineering, Muria Kudus University, Indonesia

Article Info

ABSTRACT

Impulsive spikes often occur in audio recording of gamelan where most existing methods reduce it. This research offers new method to enhance audio impulsive spike in gamelan music that is able to reduce, eliminate and even strengthen spikes. The process separates audio components into harmonics and percussive components. Percussion component is set to rise or lowered, and the results of the process combined with harmonic components again. This study proposes a new method that allows reducing, eliminating and even amplifying the spike. From the similarity test using the Cosine Distance method, it is seen that spike enhancement through Harmonic Percussive Source Separation (HPSS) has an average Cosine Distance value of 0.0004 or similar to its original, while Mean Square Error (MSE) has an average value of 0.0004 that is very small in average error and also very similar. From the Perceptual Evaluation of Audio Quality (PEAQ) testing with Harmonic Percussive Source Separation (HPSS), it has a better quality with an average Objective Difference Grade (ODG) of -0.24 or Imperceptible.

Corresponding Author:

Solekhan,
Department of Electrical Engineering,
Institut Teknologi Sepuluh Nopember,
Muria Kudus University, Kudus, Indonesia.
Email: solekhan@umk.ac.id

1. INTRODUCTION

Indonesia has many islands and thus, has diversity of tribes, customs, language, and culture. Consequently, it affects the existence of variety of musical instruments, such as gamelan. Gamelan is a musical instrument in Indonesia especially in Java and Bali [1]. It has different type and standard of notation compared to Western musical instruments.

In general, the central Javanese gamelan can be grouped into the gongan, balungan, penerusan and kendang. Gongan is a group that consists of kempyang, kempul, kenong, kethuk, and gong. Balungan is composed of saron panerus, saron barung, demung, slenthem and slentho, while panerusan consists of Bonang, gender, gambang, siter, rebab and celempung, suling, and kendang group consists of kendang and bedug. Most of the gamelan musical instruments consist of percussive music, that some materials of gamelan are made of metal, brass, silver, wood, and leather. To play the instrument is performed by hitting it. Materials used to hit usually made from wood and horn.

On the audio recording of gamelan, there are frequent impulsive spikes as shown in Figure 1. Several methods have been proposed to address the spike impulsive noise [2], [3], reducing through the envelope method [4], [5] and through the median filter method [3], [6], in order to eliminate or reduce the spike impulsive noise. The median methods are good enough for the reduction, but this method applies on all data signals without considering the harmonic or Percussive components.

In research conducted by Amart Sulong, Teddy Surya Gunawan, Othman O. Khalifa, Mira Kartiwi, Hassan Dao [7] for Speech Audio Signal Enhancement and Kayode Francis Akingbade, Isiaka A jewale Alimi [8] that explained the process of separation on audio music using Least-Mean-Square (LMS) Adaptive Algorithm and J. Driedger, M. Müller and S. Ewert [9] and H. Tachibana, N. Ono and S. Sagayama [10] that explained the process of separation between the harmonic and percussive components from the audio signal. Spike or impulsiveness tends to occur only in the percussive component.
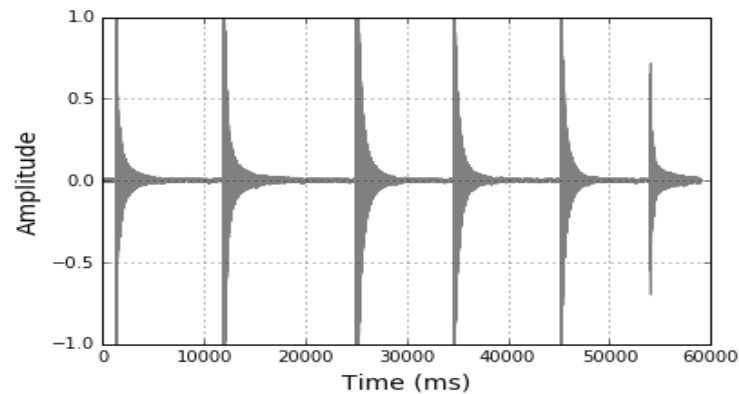

Figure 1. Spikes on audio

This research proposes the process of spike enhancement using HPSS, while for the comparison spike enhancement method is using median filter method. This paper is organized as follows. The research method in section 2, the results and discussion are in section 3 and the conclusion of the paper is in section 4.

2.   RESEARCH METHOD
2.1. Spike enhancement
The enhancement of the spike is process to enhance spike signals by reducing or amplifying. Some spike noise filters can be used to improve quality and reduce noise from signal interference. The median filter is a method that can be used to enhance by reducing the presence of spike signals. This method replaces data signals by utilizing the median average, from all data in a range.

The median is taken from a set of samples that is derived by sorting the sample in ascending or descending order, and then choose the center values. In the median filtering, a window that had been arranged sequentially over the sign and the middle of the sample in the window will be replaced by the median of all the samples in the window [6], as illustrated in Figure 2.
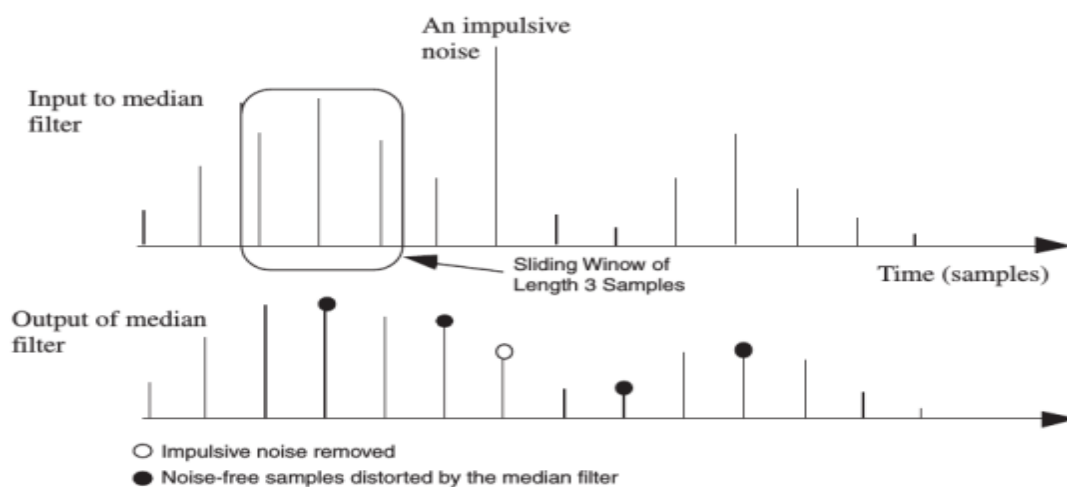

Figure 2. Input and output median filter [6]

In some experiments it showed that the median filter can be used to reduce the Spike signals well. The Median Filter output $x_{med}(n)$ with input $x(n)$ and the median window of length $2K+1$ sample data can be written as in (1),

$$x_{med}(n) = median[x(K-n), ..., x(n), ..., x(K+n)] \qquad (1)$$

Median is the middle value of a given data dimension one, from (1), n is the discrete time index, x is the signal source, K is integer number and x_{med} is the result median filter.

The purpose of the Percussive Harmonic Source Separation (HPSS) is to separate the input signal audio into a signal audio which contains the harmonic and percussive component [6]. The algorithm is based on the fact that audio has a harmonic and percussive component. The harmonic audio component tends to form the horizontal structure while the percussive audio component tends to form a vertical structure. In general, sounds can be classified into two groups: harmonic and percussive sounds. Harmonic sound is as pitched sound or melody while percussive sound is like temporary sound of percussive music instruments.

From separation process, the harmonic and percussive signal components are obtained. Spikes are the impulsive component of percussive musical audio, so by using harmonic and percussive component separation, it would be easier to enhance (by reducing, amplifying or removing) the spike component. The enhancement of the spike done in percussive component and then the result will be combined with harmonic component in the output. In some conditions, if necessary, to eliminate the percussive component, it can be more easily done by taking the harmonic components only.

2.2. Cosine similarity

Cosine similarity (CS) [11], [12] are measures based on the angle of the signal. Two signals of x and $\hat{x}$, with the cosine similarity, is described as the (2). Where CS is the Cosine Similarity, x is the source signal, and $\hat{x}$ is the new signal of x, the ranges of similarities of -1 which means the opposite, and of 1 which means equal, 0 usually indicates free. Cosine similarity (CS) can be seen as a method of normalizing document length during comparison.

$$CS = \frac{x \cdot \hat{x}}{\|x\|\|\hat{x}\|} = \frac{\sum_{n=1}^{N} x(n)\,\hat{x}(n)}{\sqrt{\sum_{n=1}^{N} x(n)^2}\,\sqrt{\sum_{n=1}^{N} \hat{x}(n)^2}} \qquad (2)$$

Cosine Distance (*CD*) is a development of the cosine similarity with relationships such as (3). This test is also used for the similarity between two vectors of audio music signals [12]. Cosine Distance (*CD*) ranges from 0 to 2 while the zero value means exactly the same and two means the opposite.

$$CD = 1 - CS \qquad (3)$$

2.3. Noise ratio

The mean squared error (*MSE*) is measures the squares average of the errors or deviations, as in (4). The MSE is a measure of the quality of an estimator, it is always non-negative, and values closer to zero are better [14]. Where *MSE* is the Mean Square Error, m is the discrete time index, N is the length of data, x is the source signal, and $\hat{x}$ is the new signal of x after enhance spike.

$$MSE = \frac{1}{N} \sum_{n=0}^{N} \left(x(m) - (\hat{x})(m) \right)^2 \qquad (4)$$

2.4. Perceptual evaluation of audio quality (PEAQ)

Perceptual Evaluation of Audio Quality (PEAQ) is standard that describe the testing audio quality objectively. This parameter is compared to the difference in the audio sound by measuring the original audio data and audio data to be tested. PEAQ Parameter can be seen on the scale of measurement in Table 7, based on ITU-R BS. 1387 [15].

The *Perceptual Evaluation of Audio Quality* (PEAQ) is estimated by mapping Model Output Variables (MOV), to single number using an artificial neural network structure with one hidden layer. The activation function of the neural network is an asymmetric sigmoid. The inputs are mapped to a *distortion index (DI)*. From (5), the network uses *I* inputs and *J* nodes in the hidden layer. The mapping is

defined by a set of input scaling factors $a_{min}[i]$, $a_{max}[i]$, a set of input weights $w_{in}[i]$, a set of output weights $w_{out}[j]$ and a pair of output scaling factors b_{min} and b_{max}.

$$DI = w_{out}[J] + \sum_{j=0}^{J-1}\left(w_{out}[j] \cdot \frac{1}{1 + e^{-\left(w_{in}[I,j] + \sum_{i=0}^{I-1} w_{in}[i,j] \cdot \frac{MOV[i] - a_{min}[i]}{a_{max}[i] - a_{min}[i]}\right)}}\right) \tag{5}$$

For the Basic version, Model Output Variables (MOV) parameters are input to a neural network with 11 input nodes, 1 hidden layer with 3 nodes and a single output. While in the Advanced version, the scaled and shifted MOV's are input to a neural network with 5 input nodes, 1 hidden layer with 5 nodes and a single output, where I is the number of MOV's (11 for the Basic version, 5 for the Advanced version) and J is the number of nodes in the hidden layer.

Model Output Variables in Basic Version uses only the FFT-based ear model. It uses the following 11 MOV's. These 11 MOV's are mapped to a single quality index using a neural network with three nodes in the hidden layer. The parameters of the mapping are given in Tables 1 to 3.

Table 1. Model Output Variables used and Scaling Factors for the Inputs of the Basic Version [15]

Index (i)	Purpose	MOV	a_{min}	a_{max}
0	Linear distortions Reference	BandwidthRef$_B$	393.916656	921
1	Linear distortions Test	BandwidthTest$_B$	361.965332	881.131226
2	Total Noise-to-mask ratio	Total NMR$_B$	−24.045116	16.212030
3	Changes in modulation	WinModDiff1$_B$	1.110661	107.137772
4	Detection probability	ADB$_B$	−0.206623	2.886017
5	Harmonic structure of the error	EHS$_B$	0.074318	13.933351
6	Average Changes in modulation	AvgModDiff1$_B$	1.113683	63.257874
7	Average Changes in modulation	AvgModDiff2$_B$	0.950345	1145.018555
8	Loudness of the distortion	RmsNoiseLoud$_B$	0.029985	14.819740
9	Detection probability	MFPD$_B$	0.000101	1
10	Frequency of audible distortions	RelDistFrames$_B$	0	1

Weights for the input nodes w_{in} in basic versions can be seen in Table 2, Weights for the output nodes w_{out} in basic versions can be seen in Table 3.

Table 2. Weights for the Input Nodes of the Basic Version [15]

Index i	MOV	Weight $w_{in}[i,0]$	Weight $w_{in}[i,1]$	Weight $w_{in}[i,2]$
0	BandwidthRef$_B$	−0.502657	0.436333	1.219602
1	BandwidthTest$_B$	4.307481	3.246017	1.123743
2	Total NMR$_B$	4.984241	−2.211189	−0.192096
3	WinModDiff1$_B$	0.051056	−1.762424	4.331315
4	ADB$_B$	2.321580	1.789971	−0.754560
5	EHS$_B$	−5.303901	−3.452257	−10.814982
6	AvgModDiff1$_B$	2.730991	−6.111805	1.519223
7	AvgModDiff2$_B$	0.624950	−1.331523	−5.955151
8	RmsNoiseLoud$_B$	3.102889	0.871260	−5.922878
9	MFPD$_B$	−1.051468	−0.939882	−0.142913
10	RelDistFrames$_B$	−1.804679	−0.503610	−0.620456
11	Bias	−2.518254	0.654841	−2.207228

Table 3. Weights for the Output Node of the Basic Version [15]

Weight $w_{out}[0]$	Weight $w_{out}[1]$	Weight $w_{out}[2]$	Bias $w_{out}[3]$
−3.817048	4.107138	4.629582	− 0.307594

Model Output Variables (MOV) in *Advanced Version* uses both the filter bank-based ear model and the FFT-based ear model. It uses the 5 MOV's. These 5 MOV's are mapped to a single quality index using a neural network with five nodes in the hidden layer. The parameters of the mapping are given in Tables 4 to 6.

Table 4. Model Output Variables used and scaling factors for the input nodes of the Advanced Version [15]

index i	MOV	Purpose	min	max
0	RmsModDiff$_A$	Changes in modulation (related to roughness)	13.298751	2166.5
1	RmsNoiseLoudAsym$_A$	Loudness of the distortion	0.041073	13.24326
2	Segmental NMR$_B$	Noise-to-mask ratio	−25.018791	13.46708
3	EHS$_B$	Harmonic structure of the error	0.061560	10.226771
4	AvgLinDist$_A$	Linear distortions (frequency response etc.)	0.024523	14.224874

Weights for the input nodes in basic version can be seen in Table 5, and weights for the output nodes in advance version can be seen in Table 6.

Table 5. Weights for the Inputs of the Advanced Version [15]

Index i	MOV[i]	Weight $w_{in}[i,0]$	Weight $w_{in}[i,1]$	Weight $w_{in}[i,2]$	Weight $w_{in}[i,3]$	Weight $w_{in}[i,4]$
0	RmsModDiff$_A$	21.211773	−39.913052	−1.382553	−14.545348	−0.320899
1	RmsNoiseLoudAsym$_A$	−8.981803	19.956049	0.935389	−1.686586	−3.238586
2	Segmental NMR$_B$	1.633830	−2.877505	−7.442935	5.606502	−1.783120
3	EHS$_B$	6.103821	19.587435	−0.240284	1.088213	−0.511314
4	AvgLinDist$_A$	11.556344	3.892028	9.720441	−3.287205	−11.031250
5	bias	1.330890	2.686103	2.096598	−1.327851	3.087055

Table 6. Weights for the Output Node of the Advanced Version [15]

$w_{out}[i,0])$	$w_{out}[i,1])$	$w_{out}[i,2])$	$w_{out}[i,3])$	$w_{out}[i,4]$	Bias $w_{out}[5]$
−4.696996	−3.289959	7.004782	6.651897	4.009144	−1.360308

Equation objective difference grade (ODG) can be seen in the equation 6. Where b_{min} and b_{max} is the scaling factor for output, see Table 7,

$$ODG = b_{min} + (b_{max} - b_{min}) \cdot \frac{1}{1 + e^{-(DI)}} \tag{6}$$

The b_{min} value is the setting minimum output value (-3.98), and the b_{max} value is setting maximum output values (0.22), so the equation becomes,

$$ODG = -3.98 + (0.22 + 3.98) \cdot \frac{1}{1 + e^{-(DI)}} \tag{7}$$

The value of the ODG was obtained from the calculation of the Model Output Variables (MOV) through neural network. As shown in Table 7.

Table 7. ObjectiveDifference Grade (ODG) [15]

Difference Grade	Description of Impairments
0	Imperceptible
-1	Perceptible but not annoying
-2	Slightly annoying
-3	Annoying
-4	Very annoying

2.5. Enhance spike using HPSS method

In this process the input audio signal is first converted using the STFT, then from the pattern of the spectrogram (Y) a median filtering of horizontal direction is done to get (Y_h) and for the vertical direction (Y_p). From the pattern, next is determining the harmonic masking (M_h) and percussive masking (M_p). The enhance spike it uses HPSS as discribed in Figure 3.
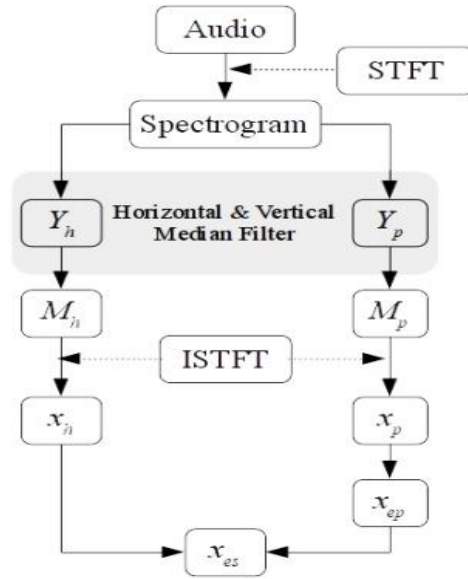
Figure 3. Enhancement spike using HPSS [5]

The results of harmonic and percussive masking is, through inversion of STFT can be obtained harmonic signals (x_h) and percussive signals (x_p). Then these percussive signals are reduced (x_{ep}) and merged back with the harmonic components (x_{es}).

Signal x is assumed to contain harmonic components (x_h) and percussive components (x_p) as in (8),

$$x = x_h + x_p \tag{8}$$

To do the separation, at the first step, compute the short-time Fourier transform (STFT) X of the signal x as shown in (9),

$$X(m,k) = \sum_{n=0}^{N-1} x\,(n+mH)\omega(n)e^{(-2\pi ikn/N)} \tag{9}$$

With $m \in [0{:}M-1]$ and $k \in [0{:}N-1]$, M the number of frames, N the size of the frame, $\omega \in [0{:}N-1] \rightarrow R$ is a function of window and H is the hop size of X can be searched the spectrogram Y such as (10).

$$Y(m,k) = \left(X(m,k)\right)^2 \tag{10}$$

Next process is applying median filter to spectrogram both horizontally and vertically as in (11) and (12) to get the horizontal spectrogram component Y_h and vertical spectrogram component Y_p.

$$Y_h = Y_{medh}(m,k) = median[Y(K-m,k),\dots,Y(m,k),\dots,Y(K+m,k)] \tag{11}$$

$$Y_p = Y_{medp}(m,k) = median[Y(m,K-k),\dots,Y(m,k),\dots,Y(m,K+k)] \tag{12}$$

With Yh is the horizontal Spectrogram from spectrogram Y, Yp is the vertical spectrogram of spectrogram Y, m is the data on horizontal, k is the data on vertical, and K in integer value. The binary masking can be obtained with equations (13) and (14).

$$M_h(m,k) = \begin{cases} 1 & Y_h(m,k) \geq Y_p(m,k) \\ 0 & other \end{cases}$$ (13)

$$M_p(m,k) = \begin{cases} 1 & Y_p(m,k) \geq Y_h(m,k) \\ 0 & other \end{cases}$$ (14)

M_h and M_p are horizontal and vertical binary masking respectively. Then, to get harmonic and percussive component can be done by using Equations (15) and (16)

$$X_h = X(m,k)M_h(m,k)$$ (15)

$$X_p = X(m,k)M_p(m,k)$$ (16)

X_h is harmonic component in frequency domain, while X_p is the percussive component. By performing the inverse STFT, then harmonics and percussive components can be obtained by Equations (17) and (18).

$$x_h(m) = \frac{1}{N}\sum_{k=0}^{N-1} X_h(m,k)e^{j2\pi km/N}$$ (17)

$$x_p(m) = \frac{1}{N}\sum_{k=0}^{N-1} X_p(m,k)e^{j2\pi km/N}$$ (18)

With x_p percussive component and x_h is a harmonic components of signal x, the next percussive can be reduced as x_{ep}, as shown in (19).

$$x_{ep} = ef\, x_p$$ (19)

With *ef* as enhance factor, in the (20), a new signal is defined by the combination between results of reduction and harmonic components as x_{es}.

$$x_{es} = x_h + x_{ep}$$ (20)

The result x_{es} is a combination of a new signal in which Percussive component has been enhanced.

3.   RESULTS AND DISCUSSION

The steps of enhancement spike testing using HPSS method can be explained as follows; Experiments are applied to signal as shown in Figure 4, which is a signal that contains a spike. Then, the signal is modified to shape the spectrogram patterns as shown in Figure 5. Median filter processes are implemented against the spectrogram horizontally and vertically for getting binary masking as in (8) until 14. It results harmonic and percussive component as in (15) and 16. This result can be seen in Figure 6 (Harmonic Power Spectrogram) and Figure 7 (Percussive Power Spectrogram).

The signal separation converted back using the Inverse STFT ((17) and (18)), as shown in Figure 8 and 9 and then the percussive signal components is enhanced as shown in Figure 10. The next process is to recombine the enhanced percussive signal and harmonic signal, as shown in Figure 11.
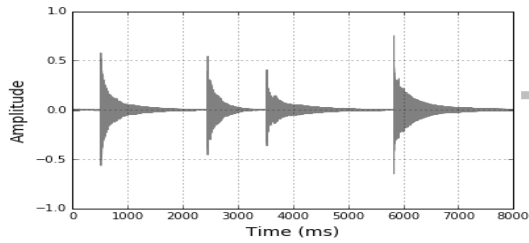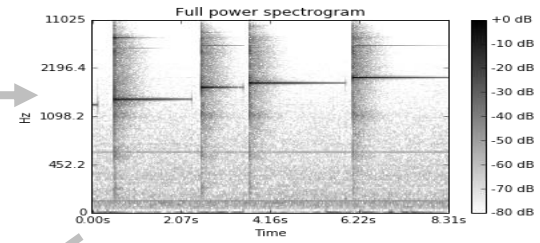
Figure 3. Source signal
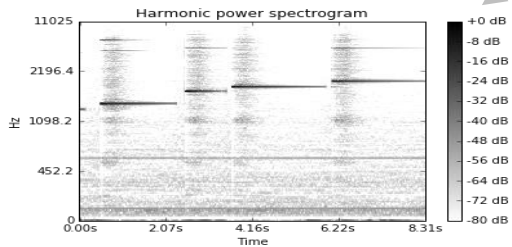

Figure 4. Source signal spectrogram
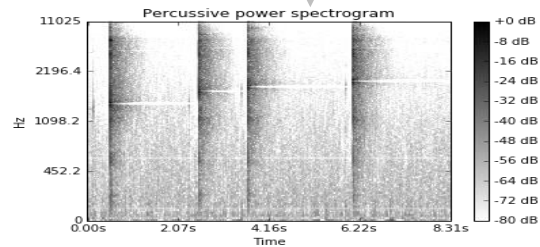

Figure 5. Harmonic power spectrogram

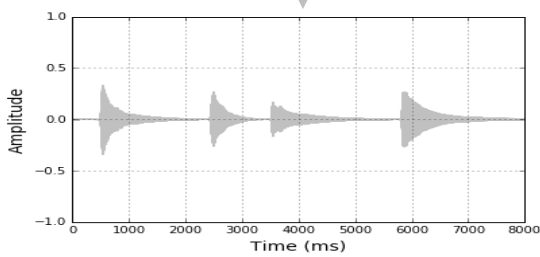Figure 6. Percussive power spectrogram
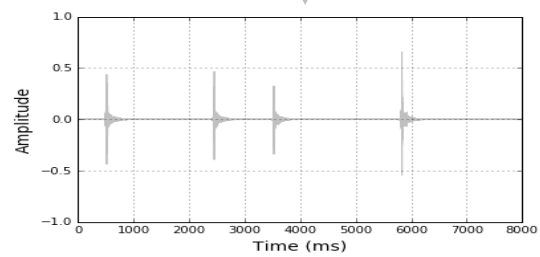

Figure 7. Harmonics component
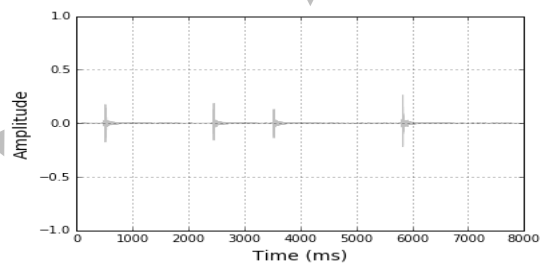

Figure 8. Percussive component

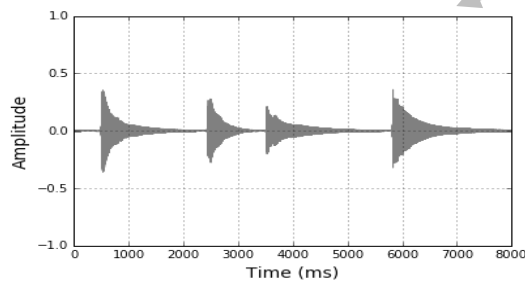Figure 9. Enhance of percussive component

Figure 10. Reconstructed signal

Audio recordings of music are taken from ITS gamelan instruments, while the experiment and testing are conducted in the B303 laboratory. This experiment then performed using 132 audio signals to examine enhance spike using median filter. According to (1), the value of K is in the range of 1 to 6.

The results of this experiment are tested by using Cosine Distance (CD), Mean Square Error (MSE) and Objective Difference Grade (ODG), for the test of spikes enhance using a median filter with index K 1 to 6 is shown in Table 8.

Table 8. Enhance Spike Using Median Filter

K	CD	MSE	ODG
1	0.0000	0.0000	-0.24
2	0.0001	0.0001	-0.82
3	0.0004	0.0006	-1.44
4	0.0009	0.0019	-1.99
5	0.0020	0.0046	-2.38
6	0.0040	0.0089	-2.67
AVG	0.0012	0.0027	-1.59

CD=Cosine Distance, MSE=Mean Square Error, ODG=Objective Difference Grade

From Table 8, it is found that the results of the enhanced spike with reduction using median filter method, the smaller range is better with the CD and MSE values close to zero. This indicates that median filters can be used for audio spike reduction of gamelan music audio.

For some circumstances, spike on some results of the audio recording of the gamelan has low magnitude, so it needs amplifying due to enhancement spike using median filter method cannot amplify spike. HPSS method can be used to solve this issue.

The experiment and testing to enhance spike using HPSS method performed using 132 audio signals, with Enhance Factor (EF) from 1.3 to 0.7. The value of enhanced factor 1 means no amplify or reduction, so it is not used; while reduction spike for EF is less than one ($EF < 1$), and amplify spike for EF is greater than one ($EF > 1$). The result of this experiment is tested with the Cosine Distance (CD), Mean Square Error (MSE) and Objective Difference Grade (ODG), using HPSS with EF 0.7 to 1.3 that can be seen in Table 9.

From Table 8 and Table 9, it is found that the results of the enhanced spike with the proposed method (HPSS), have smaller CD value with average of 0.0004, and the method of median filters has a CD value of 0.0012. MSE in HPSS methods has value very small average of 0.0004, while in the median filter methods have value of 0.0027.

Table 9. Enhance Spike using HPSS with EF From 1.3 to 0.7

No	EF	CD	MSE	ODG
1	0.7	0.0009	0.0008	-0.44
2	0.8	0.0004	0.0003	-0.19
3	0.9	0.0001	0.0001	-0.04
4	1.1	0.0001	0.0001	-0.05
5	1.2	0.0003	0.0003	-0.23
6	1.3	0.0007	0.0008	-0.51
AVG		0.0004	0.0004	-0.24

EF = Enhancement Factor

Testing of Perceptual Evaluation of Audio Quality (PEAQ) using median filters has average ODG value of -1.59 (Perceptible but not annoying), as shown in Table 8, while using HPSS the average ODG value is -0,24 (Imperceptible), as shown in Table 9. So the PEAQ test of enhances spike using HPSS method is better compared with the enhanced spike using median filter method. The enhancement spike with HPSS method can increase the weak spike and reduce the high spike, while the enhancement spike using median filter method can reduce it only.

4. CONCLUSION

The test performed on some gamelan audio shows that median filter method can be applied in enhancement spike, with reduction of impulsive spikes in gamelan music audio results, average value of Cosine Distance (CD) is 0.0012, Mean Square Error (MSE) value is 0.0027, and Objective Difference Grade

(ODG) value is -1.59 (imperceptible). While using HPSS methods results, average value of Cosine Distance (CD) is 0.0004, Mean Square Error (MSE) is 0.0004, and Objective Difference Grade (ODG) values is -0.24 (imperceptible). So HPSS method can be used to enhance the spike with the reduction or amplify modes.

ACKNOWLEDGEMENTS

REFERENCES

[1]   Sudarma, Made, Ariyani, Sri, Adiputra, Nyoman. "Aligner Tones Software Improve Measurement Precision In The Process of Balinese Gamelan," *International Journal of Electrical and Computer Engineering (IJECE)* Vol. 3, No.1, pp. 15~20, 2013.

[2]   Sarah Paige Gibson, "Neural Spike Sorting in Hardware: From Theory to Practice," University of California, Los Angeles, 2012.

[3]   S. Abdul Saleem, T. Abdul Razak, "An Effective Noise Adaptive Median Filter for Removing High Density Impulse Noises in Color Images, "*International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 6, No. 2, pp. 611~620, 2016.

[4]   Byunghwa Lee; Minsoo Hahn; Kwangki Kim; Jinsul Kim, "Efficient transient Signal Detection in Spatial Cue Based Multi-Channel Audio Coding," International *Conference on Information Science and Applications (ICISA)*, pp.1-3, 6-9 May 2014.

[5]   Solekhan, Yoyon K Suprapto, Wirawan, "Audio Spike Detection on Gamelan using Envelope Shape Pattern Analysis," *Telecommunication Computing Electronics and Control* (*TELKOMNIKA)*, Vol 14, no.4, pp 1432-1437, 2016.

[6]   Saeed V. Vaseghi, "Advanced Digital Signal Processing and noise Reduction," *Fourth Edition, John Wiley & Sons, Ltd.,* 2008.

[7]   Amart Sulong, Teddy Surya Gunawan, Othman O. Khalifa, Mira Kartiwi, Hassan Dao., "Single Channel Speech Enhancement using Wiener Filter and Compressive Sensing," *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 7, No. 4, pp. 1941~1951, 2017.

[8]   Kayode Francis Akingbade, Isiaka A jewale Alimi, "Separation of Digital Audio Signals using Least-Mean-Square (LMS) Adaptive Algorithm," *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 4, No. 4, pp. 557~560, 2014.

[9]   J. Driedger, M. Müller and S. Ewert, "Improving Time-Scale Modification of Music Signals Using Harmonic-Percussive Separation," *in IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 105-109, Jan. 2014.

[10] H. Tachibana, N. Ono and S. Sagayama, "Singing Voice Enhancement in Monaural Music Signals Based on Two-stage Harmonic/Percussive Sound Separation on Multiple Resolution Spectrograms," *in IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 228-237, Jan. 2014.

[11] Y. Vaizman, B. McFee and G. Lanckriet, "Codebook-Based Audio Feature Representation for Music Information Retrieval*," in IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1483-1493, Oct. 2014.

[12] J. Prazak and M. Bohac, "Speaker diarization of broadcast audio using automatic transcription, iVectors and cosine distance scoring*," ELMAR Proceedings*, Zadar, pp. 211-214, 2012.

[13] C. F. dos Reis, R. A. P. Altafim, Y. Gutnik, R. A. C. Altafim and A. C. Cavalcanti, "Comparative analysis between impulsive detection methods applied on partial discharge acoustic signals," *Instrumentation and Measurement Technology Conference (I2MTC) Proceedings, 2014 IEEE International, Montevideo, 2014*, pp. 418-421, 2014.

[14] Lehmann, E. L.; Casella, George "Theory of Point Estimation (2nd ed.)", *New York: Springer.* ISBN 0-387-98502-6. MR 1639875 1998.

[15] P. Kabal, "An Examination and Interpretation of ITU-R BS. 1387: Perceptual Evaluation of Audio Quality," *McGill Univ.*, pp. 1-96, 2003.

BIOGRAPHIES OF AUTHORS

Solekhan.was born in Pati 19th of May 1972 and recieved the bachelor degree in electrical engineering from Brawijaya University, Surabaya, Indonesia, in 1997, the Master of JCM in electrical engineering from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 2003. Since 1997 he has been with University of Muria Kudus as lecturer in the Electrical Engineering Department. His research interests include statistical signal processing, Data Mining, Audio Signal Processing and Traditional Music. He is currently pursuing the Ph.D. degree at Institut Teknologi Sepuluh Nopember (ITS) , Surabaya, Indonesia.

Yoyon K. Suprapto was born in Solo 25th of September 1954 and received the bachelor degree in electrical engineering from Institut Teknologi Bandung, Bandung, Indonesia in 1977, Master of computer science from The University of Missouri, Columbia, Missouri, USA in 1981 and Ph.D. degree at Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia (ITS) in 2010. He is currently a lecturer in the Department of Electrical Engineering, ITS. His current interests are Data Mining, Sound Signal Processing and Traditional Music. He is a member of IEEE, and IAENG.

Wirawan was born in Pasuruan 9th of Nopember 1963 and received Electrical Engineering degree from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia (ITS) in 1987, DEA (M.S. equivalent) degree from Ecole Supérieure en Sciences Informatiques, Université de Nice-Sophia Antipolis, France (ESSI-UNSA) in 1996 and Dr. degree from Télécom ParisTech, France (ENST) in 2003. He is currently a lecturer in the Department of Electrical Engineering, ITS, Surabaya. His research interests include various aspects of wireless sensor networks, multimedia signal processing and communications, and underwater acoustics communication and networking.