Y Christian¹, I D M B A Darmawan²

¹²Program Studi Teknik Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana, Jalan Raya Kampus Unud, Jimbaran, Badung, Bali, Indonesia

²dewabaya@unud.ac.id

Abstract. Rindik is a traditional music instrument originated from Bali that consists of 11 bamboo rods and played by a person hitting the bamboo rod with a rubber mallet in each player’s hands. Documentation of Rindik songs with automatic music transcription work is easier to do by separating it first. To overcome that challenge, two Rindik rods sound were separated using the spectral subtraction method. The noise spectrum is the spectrum of a single rod sound that needs to be muffled. The resulting audio is the other single rod sound and vice versa to get both single rod sound. The data consisted of single rod hit sound recordings of 11 single rod sound and 55 combinations of two-rod sounds being hit at the same time. The performance of the spectral subtraction method in separating Rindik sound was measured with MSE and SIR and also heard the noise exists in the separated audio signal. The experiment demonstrated a state-of-the-art performance consists of spectral subtraction with squared noise average magnitude with average MSE value 0.0126 and average SIR value 55.68 dB.

1. Introduction

The many traditional musical instruments originated from Bali is one of the reasons for Bali to be well known for its rich arts and culture. Rindik is a vibraphone like music instrument originated from Bali that consists of 11 bamboo rods. Rindik is played with a rubber mallet in each player’s hand. There has been little effort to document Rindik songs recently. The efforts that have been done before were documenting Rindik songs in a video or audio recording. Hence, storing the documentation in a video or audio form requires much more storage than documentation in text transcription form. Meanwhile, music transcription works are easier to do when the audio recording is separated [1]. Rindik music is played with two hands. In a Rindik song, often, two Rindik rods are simultaneously hit at the same time. In this study, the audio recording of two Rindik rods sounds being hit together with the left and right hand at the same time were separated to make the music transcription work easier. The result of the separation process was each of the single rod sounds from the two-rod sound in two separate audios. The separation method used was the spectral subtraction method. In most of the previous researches, the spectral subtraction method was commonly used to enhance speech audio signals [2–5].

Speech enhancement methods were used for several tasks, such as background elimination, noise termination, and multi-speech separation (speaker separation) [2]. There is a spectral subtraction method carried out to enhance a heart sound by suppressing the noise [6]. This indicates that the spectral subtraction method could be used for a non-speech audio signal. The spectral subtraction method has been extensively studied because of its simplicity and effectiveness [7]. The principle of this method is to subtract the spectrum of the noise from the noisy speech audio signal [7].
The separation task is commonly done to separate a vocal in a music instrument [8,9]. The separation task is also done to separate different music sources into a single instrument [10]. That separation case is not the same as the separation case in this research. The separation task to do for the Rindik music instrument is a separation method that separates two tones/two-rods sound into a single individual rod sound.

In this research, the two-rods sound was separated into two separate audio of single rod sound by subtracting the two-rods sound with the noise spectrum average magnitude in which the noise is the single rod sound. The spectral subtraction method was repeated twice in each separation process. The first one applies the method with a single rod sound that hit with the player’s left hand as noise, producing the single rod sound that hit with the player’s right hand. The second one applies the method with a single rod sound that hit with the player’s right hand as noise, producing the single rod sound that hit with the player’s left hand. The sound produced by the left hand is the Rindik rod sound which is the earlier one when it counted from the left side of the Rindik and the right hand is vice versa. It is restricted to hand-crossed hit the Rindik rod.

The performance of the separation method was measured by calculating the MSE (mean square error) and the SIR (signal to interference ratio) of the separated audio with the original audio [11]. In this case, the original audio was the original single rod sound. The higher the SIR value and the lower the MSE value, the better the result of the separation [11]. The experiment was conducted to find the best frame length. Also, some of the results still contained both of the rods sounds, so the average magnitude of the noise is squared, and the results were compared to the non-squared average magnitude of the noise.

2. Method

There are several steps to do the separation with the spectral subtraction method. The steps taken to perform the separation was begin with making a model of the spectrum average magnitude of every single rod sounds as noise and stored in the database. And then, the sound signal frames were converted from the time domain into the frequency domain. The spectrum of each frame was subtracted with the noise spectrum average magnitude model, and the audio signal in the frequency domain was reconstructed into the time domain. The figure below shows the steps of the method.

![Spectral subtraction steps in separation](image)

**Figure 1.** Spectral subtraction steps in separation.

---

The separation task is commonly done to separate a vocal in a music instrument [8,9]. The separation task is also done to separate different music sources into a single instrument [10]. That separation case is not the same as the separation case in this research. The separation task to do for the Rindik music instrument is a separation method that separates two tones/two-rods sound into a single individual rod sound.

In this research, the two-rods sound was separated into two separate audio of single rod sound by subtracting the two-rods sound with the noise spectrum average magnitude in which the noise is the single rod sound. The spectral subtraction method was repeated twice in each separation process. The first one applies the method with a single rod sound that hit with the player’s left hand as noise, producing the single rod sound that hit with the player’s right hand. The second one applies the method with a single rod sound that hit with the player’s right hand as noise, producing the single rod sound that hit with the player’s left hand. The sound produced by the left hand is the Rindik rod sound which is the earlier one when it counted from the left side of the Rindik and the right hand is vice versa. It is restricted to hand-crossed hit the Rindik rod.

The performance of the separation method was measured by calculating the MSE (mean square error) and the SIR (signal to interference ratio) of the separated audio with the original audio [11]. In this case, the original audio was the original single rod sound. The higher the SIR value and the lower the MSE value, the better the result of the separation [11]. The experiment was conducted to find the best frame length. Also, some of the results still contained both of the rods sounds, so the average magnitude of the noise is squared, and the results were compared to the non-squared average magnitude of the noise.

2. Method

There are several steps to do the separation with the spectral subtraction method. The steps taken to perform the separation was begin with making a model of the spectrum average magnitude of every single rod sounds as noise and stored in the database. And then, the sound signal frames were converted from the time domain into the frequency domain. The spectrum of each frame was subtracted with the noise spectrum average magnitude model, and the audio signal in the frequency domain was reconstructed into the time domain. The figure below shows the steps of the method.

![Spectral subtraction steps in separation](image)

**Figure 1.** Spectral subtraction steps in separation.
The steps in figure 1 were carried out twice for the two-rods sounds. Assume the rods that being hit are the third and the fifth-rods, then the subtraction method was carried out two times with the third-rods spectrum average magnitude as the noise, leaving only the fifth-rods sound and vice versa.

a. Characteristics of Rindik

*Rindik* is made of bamboo, which consists of several rods of bamboo. Bamboo rods are arranged in a row. They are tied and framed with wood. The tied bamboos then are painted and decorated with carvings. Figure 2 is the image of the *Rindik* instrument.

![Figure 2. Rindik instrument.](image_url)

There are two types of *Rindik* tones [12]. They are *lanang*, which the rhythm of a song, producing bass with longer and larger bamboo rods [12]. The other one is *wadon*, which the melody of a song, producing higher pitch sound with shorter and smaller bamboo rods [12]. They are all arranged sequentially from low to high tones from *lanang* to *wadon* [12]. Tones produced by *Rindik* are based on the *selendro* tone system consisting of only five types of tones: *nding*, *ndong*, *ndeng*, *ndung*, and *ndang* [12]. Figure 3 and 4 shows the spectrum of a *Rindik* rod sound analyzed with Audacity software.

![Figure 3. The spectrum of fifth *Rindik* rod.](image_url)

![Figure 4. The spectrum of fifth and tenth *Rindik* rods sound.](image_url)

b. Dataset

The data were observed by recording directly from one *Rindik* music instrument sound source. The data were recorded from one *Rindik* instrument that has 11 bamboo rods, played by one person. The collected data were 11 single rod sound and 55 combinations of two-rods sound that being hit at the same time recorded in wave (.wav) file format. All of the audio recording files have a 44100 sampling rate with 16
bits encoding. The microphone of the recorder was placed on the left side of the Rindik instrument while recording. Figure 5 shows how the sound of Rindik rods was recorded into the dataset.

![Figure 5. Rindik sound recording set up.](image)

c. Spectral Subtraction
Spectral subtraction is a method used to restore the power or magnitude spectrum of a signal observed in additive noise by subtraction of the average noise spectrum from the noisy signal spectrum [13]. The noise spectrum was usually calculated when there are no sound signals, and only noise signals are present [13]. The computational complexity of the spectral subtraction method is relatively inexpensive [13]. In performing spectral subtraction on the spectrum, this method can produce negative values on the subtracted spectrum, while the power or magnitude spectrum is a non-negative variable [13]. Hence, negative results must be converted into non-negative values [13]. The process of converting negative values to non-negative values is half-wave rectification [4].

The spectral subtraction method was started with segmenting the audio signal in frames with half-frame blocking hop size and the frames windowed with the Hamming window [4]. The windowed frames were transformed with the fast Fourier transform (FFT). The two-rods sound frames spectrum was subtracted with the single rod average magnitude and applies half-wave rectification to replace the negative values with zero. The reconstruction of the audio signal used the inverse fast Fourier transform. The result was evaluated by calculating the mean square error and the signal to interference ratio [11].

d. Frequency Domain Transformation with FFT
The variation of the frame length used are 1024, 2048, and 4096 samples per frame. Each frame was multiplied with the Hamming window function (1) as follows [14].

$$w(n) = 0.54 - 0.64 \cos \cos \left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq N - 1$$

$N$ is the number of Hamming window sample and the frame length. The data stored in the Hamming window were calculated for spectral calculations, which is a fast Fourier transform calculation. The general equation for the discrete Fourier transform (DFT) is as follows (2).
\[ \tilde{X}[k] = \sum_{n=0}^{N-1} \tilde{X}[n] e^{-j(2\pi k/N)n} \] (2)

The DFT is a function of \( k \) in the frequency domain. Each function of \( k \) is an impulse function with a magnitude, frequency, and phase shift. The FFT is an algorithm that efficiently computes the DFT. The FFT of each time-windowed data buffer is called a time frame [5].

e. Spectral Subtraction and Frequency-Time Domain Transformation with IFFT

The calculation for the spectrum average magnitude of the noise was done by summing each frequency component in the spectrum and divided by the number of frames (3). Hence, the average magnitude of the noise spectrum was obtained and will be used in the subtraction process. The equation for calculating the average magnitude is as follows [4].

\[ \mu(k) = E[|N[k]|] \] (3)

Since the subtracted spectrum still contained both of the sounds of the rod, this research tried to square the average magnitude (4) of the noise spectrum to make the frequency component of the noise reduced with the equation as follows.

\[ \mu(k) = E[|N[k]|]^2 \] (4)

\( E \) is the average value operator. Each frame of the audio signal was subtracted with the average magnitude of the noise. Spectral subtraction (5) of each frame was conducted with the following equation [4].

\[ \tilde{S}[k] = |\tilde{X}[k]| - \mu(k) \] (5)

The frequency spectrum is a non-negative value. After the subtraction process was carried out, there may be a frequency magnitude with a negative value in the spectrum so that those negative values were replaced with zero by using half-wave rectification (6). The equation for this operation is as follows [4].

\[ \hat{S}[k] = \begin{cases} 0, & \tilde{S}[k] < \mu(k) \\ \tilde{S}[k], & \tilde{S}[k] \geq \mu(k) \end{cases} \] (6)

Transformation of the subtracted spectrum from the frequency domain into the time domain was carried out with the inverse FFT method (7) with the equation as follows [13].

\[ \hat{S}[n] = \frac{1}{N} \sum_{k=1}^{N} \hat{S}[k] e^{j\theta_V(k)} e^{j(2\pi k/N)n}, \quad 0 \leq n \leq N - 1 \] (7)

Where \( \theta_V(k) \) is the phase of the audio signal frequency \( Y(k) \) [13].

f. Evaluation

Evaluation of spectral subtraction method in the separation process was calculated with the mean square error and the signal to interference ratio in decibel unit (dB). The evaluation was not only calculate the MSE and the SIR; the audio result of the spectral subtraction method were listened to whether there was still noise or clean from the noise.
MSE was a method for measuring the difference between the estimator (reconstruction signal) and the actual value (baseline signal) [11]. The actual value used in the evaluation was the original single-rod sound. The evaluation consisted of calculating the MSE and the SIR of the separated audio with the original single-rod audio. The mean square error (8) was calculated with the equation as follows [15].

\[
MSE = \frac{1}{N} \sum_{i} (X_i - Y_i)^2
\]  

(8)

\(X_i\) is the original signal, and \(Y_i\) is the result (8) of the subtracted signal. SIR is the ratio of a signal to interference and commonly used to measure the quality of a signal to interference [11]. The higher the SIR value, the better the quality of the signal [11]. The signal to interference ratio (9) was calculated with the equation as follows [11].

\[
SIR = -10 \log(MSE)
\]  

(9)

Where MSE is the variable of the mean square error.

3. Result and Discussion

The spectral subtraction process implemented with Python programming language. Fifty-five combinations of two Rindik rods sound and 11 single rod sounds were recorded for the experiment. All of the two Rindik rods sounds were subtracted twice, with each of the single rod sounds that existed in the two Rindik rods audio signal.

a. Variation of The Frame Length

The experiment was carried out with several variations of frame length, which are 1024, 2048, and 4096 samples. The frame length with the highest SIR and the lowest MSE value was the best in separating Rindik rod sound.

| Frame Length (sample) | Average SIR (dB) | Average MSE |
|-----------------------|------------------|-------------|
| 1024                  | 51.74            | 0.0125      |
| 2048                  | 51.85            | 0.0127      |
| 4096                  | 51.98            | 0.0130      |

Table 1 shows that the best frame length used in the spectral subtraction process is the frame with 4096 samples length. The frame with 4096 samples resulted in the highest average SIR, which is 51.98 dB, but not the lowest MSE. However, the 4096 sample frame length has better audio results if it was listened to.

b. Squared and Non-Squared Average Magnitude

After the subtraction processes were carried out, in fact, some of the audio results still contained both of the sounds of the rod. It occurred because the average noise spectrum magnitude cannot significantly subtract the spectrum of the single rod sound from the spectrum of the two-rods sound. Thus, this research tried to square the average magnitude of the noise spectrum and resulted in the subtracted audio signals with less noise. Table 2 shows the result of the spectral subtraction with squared and non-squared average magnitude.
Table 2. Squared and non-squared average magnitude result.

| Average Magnitude | Average SIR (dB) | Average MSE |
|-------------------|-----------------|-------------|
| Squared           | 55.68           | 0.0126      |
| Non-squared       | 51.98           | 0.0130      |

The experiment of squared and non-squared average magnitude was carried out with the best frame length, which is 4096 samples length. Table 2 shows the comparison of average SIR and average MSE for the result of the spectral subtraction method in separating two-rods sound. The squared average magnitude resulted in a higher average SIR value, which is 55.68 dB, and a lower MSE value, which is 0.0126. If the audio results have listened, most of the separated audio contained a single rod sound. This result proved that the spectral subtraction can do the separation task for the Rindik music instrument.

4. Conclusion

Rindik is one of the traditional music from Bali. In this research, two Rindik rods sounds that being hit at the same time are separated into two single rod audio signals using the spectral subtraction method. After experimenting, a few audio signal results still contained the noise so that the average magnitude of the noise spectrum was squared, and the noise can be subtracted significantly from the spectrum. Most of the audio results only contained a single rod sound. The variation of frame length and experiment in squaring the average magnitude showed the best frame length is 4096 samples length with the squared average magnitude resulting in average SIR 55.68 dB and average MSE 0.0126.

References

[1] Klapuri A and Davy M 2006 Signal processing methods for music transcription (New York: Springer)
[2] Jain R R and Rajtank V 2018 Implementation of Spectral Subtraction Using IBM in Simulink 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017
[3] Saldanha J C and Shruthi O R 2017 Reduction of noise for speech signal enhancement using Spectral Subtraction method Proceedings - 2016 International Conference on Information Science, ICIS 2016
[4] Karam M, Khazaal H F, Aglan H and Cole C 2014 Noise Removal in Speech Processing Using Spectral Subtraction J. Signal Inf. Process.
[5] Cole C, Karam M and Aglan H 2008 Spectral subtraction of noise in speech processing applications Proceedings of the Annual Southeastern Symposium on System Theory
[6] Zhang L and Wang L 2016 Heart Sound Enhancement Based on Improved Spectral Subtraction Proceedings - 11th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2015
[7] Li C and Liu W J 2011 A novel multi-band spectral subtraction method based on phase modification and magnitude compensation ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings
[8] Jansson A, Humphrey E, Montecchio N, Bittner R, Kumar A and Weyde T 2017 Singing voice separation with deep U-Net convolutional networks Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017
[9] Kumar R, Luo Y and Mesgarani N 2018 Music source activity detection and separation using deep attractor network Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH
[10] Park S, Kim T, Lee K and Kwak N 2018 Music source separation using stacked hourglass networks Proceedings of the 19th International Society for Music Information Retrieval
Conference, ISMIR 2018

[11] Putra A P, Wiantari N W, Dewi N P M N and Darmawan I D M B A 2019 Independent Component Analysis (ICA) dan Sparse Component Analysis (SCA) dalam Pemisahan Vokal dan Instrumen pada Seni Geguntangan J. Elektron. Ilmu Komput. Udayana 8 105–11

[12] Widiartha I M and Muliantara A 2017 Rindik Voice Synthesis Using Modified Frequency Modulation as Bali Cultural Preservation Efforts Kursor

[13] Vaseghi S V. 2000 Spectral Subtraction Advanced Digital Signal Processing and Noise Reduction (New Jersey: John Wiley & Sons Ltd) pp 333–7

[14] Hamid O K 2018 Frame Blocking and Windowing Speech Signal J. Information, Commun. Intell. Syst. 4

[15] Aggarwal R, Rathore S, Singh J K, Tiwari M, India M P, Gupta V K and Khare A 2011 Noise Reduction of Speech Signal using Wavelet Transform with Modified Universal Threshold Int. J. Comput. Appl.