Audio Spectrum Processing for Sound-based Non-destructive Methods Used in Fruit Quality Determination

Viet-Hong Tran[1,2*], Duc Thang Dao[1,2], Tuan Nghia Nguyen[1,2]

[1]Department of Mechatronics Engineering, Faculty of Mechanical Engineering, Ho Chi Minh city University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam

[2]Vietnam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Vietnam

Email address: tvhong@hcmut.edu.vn

Abstract—*Despite significant advancements in science and technology, assessing the maturity of certain fruits in large quantities, such as durian, pineapple, watermelon, coconut, and walnut, remains a challenging task. To address this, non-destructive methods are commonly employed to determine fruit maturity. One such approach involves studying the sound obtained from tapping on the fruits' husks. Depending on the presence of ambient noise during knocking, the need for noise filtering becomes apparent. In quiet environments with minimal noise, noise filtering may not be necessary. However, in noisy environments, a noise reduction algorithm becomes crucial to enhance the intrinsic features of the fruit's sound while attenuating external contaminating components. In this research, we propose a noise reduction algorithm for the aforementioned purpose.*

Keywords— *Noise reduction, Audio spectrum processing, Fruit quality determination.*

I. INTRODUCTION

The durian export industry has become a significant source of economic benefits for Vietnam, owing to the fruit's distinct flavour and creamy texture, making it highly popular worldwide. However, determining the ideal ripeness of durian poses a challenge, as it is difficult to do so visually. Harvesting the fruit too early results in a lack of flavour, while harvesting it too late may not meet economic demands. Many studies focusing on the sound characteristics of durian knocking as a non-destructive method for ripeness classification have been undertaken. However, recording durian knocking in the garden introduces environmental disturbances such as wind-caused leaf sounds, human chatter, vehicle noise, and street sirens.

Presently, the majority of research focused on determining fruit maturity tends to overlook the noise reduction methodology, often opting for controlled experimental settings with minimal disruptions. However, this approach proves impractical in real-world scenarios as it disregards the practical constraints faced by farmers and companies engaged in fruit production, the eventual beneficiaries of such research findings. These prospect users find this methodology financially burdensome. Therefore, it is necessary to utilize noise reduction algorithm in such scenario. However, some algorithms could overestimate noise, which results in the alteration of the characteristics of the original audio signal in the process. Consequently, those algorithms introduce distortions and the loss of main features into the recording. Thus, the chosen algorithm should cause minimal or no distortion to the audio recording.

The three most prevalent methods for noise reduction are spectral subtraction, Wiener filtering, and Kalman filtering [1]. Wiener filtering [1], a technique that employs a linear filter with consistent properties over time, aims to estimate a desired random process by eliminating introduced noise, leading to a reduction in signal's mean square error. Kalman filtering [1], on the other hand, employs a recursive approach, offering rapid responses ideal for dynamic systems. It follows a two-step process, initially generating estimates for present state variables along with their associated uncertainties. As signals often contend with random noise, subsequent measurements prompt adjustments to these estimates through a weighted average mechanism, giving priority to more uncertain estimates. The widely embraced Spectral Subtraction Algorithm [1] is chosen for its simplicity in implementation, relying on amplitude and phase of the clean signal to calculate noise elements, thus proving to be a frequently employed strategy. These algorithms all have its own limitation in the field of fruit maturity determination.

The chosen algorithm is typically applied in environments where intricate noise patterns are prevalent, particularly when faced with noise of considerable variability. When employing the Spectral Subtraction Algorithm, there's an increased possibility of unintentionally neglecting a minor fraction of the noise component, leading to the occurrence of a phenomenon termed musical noise. Meanwhile, in the context of Kalman filtering and Wiener filtering, there exists a propensity to overestimate the noise, potentially leading to the removal of crucial features from the target signal. In the forthcoming paper, we will demonstrate how our algorithm has the possibility to address these limitations, especially in the field of fruit maturity determination.

II. PROPOSED NOISE REDUCTION ALGORITHM

In this paper, one sound source is observed and only one instance of the signal of interest is considered. One microphone is used to collect the observed signal which is degraded by the noise signal.

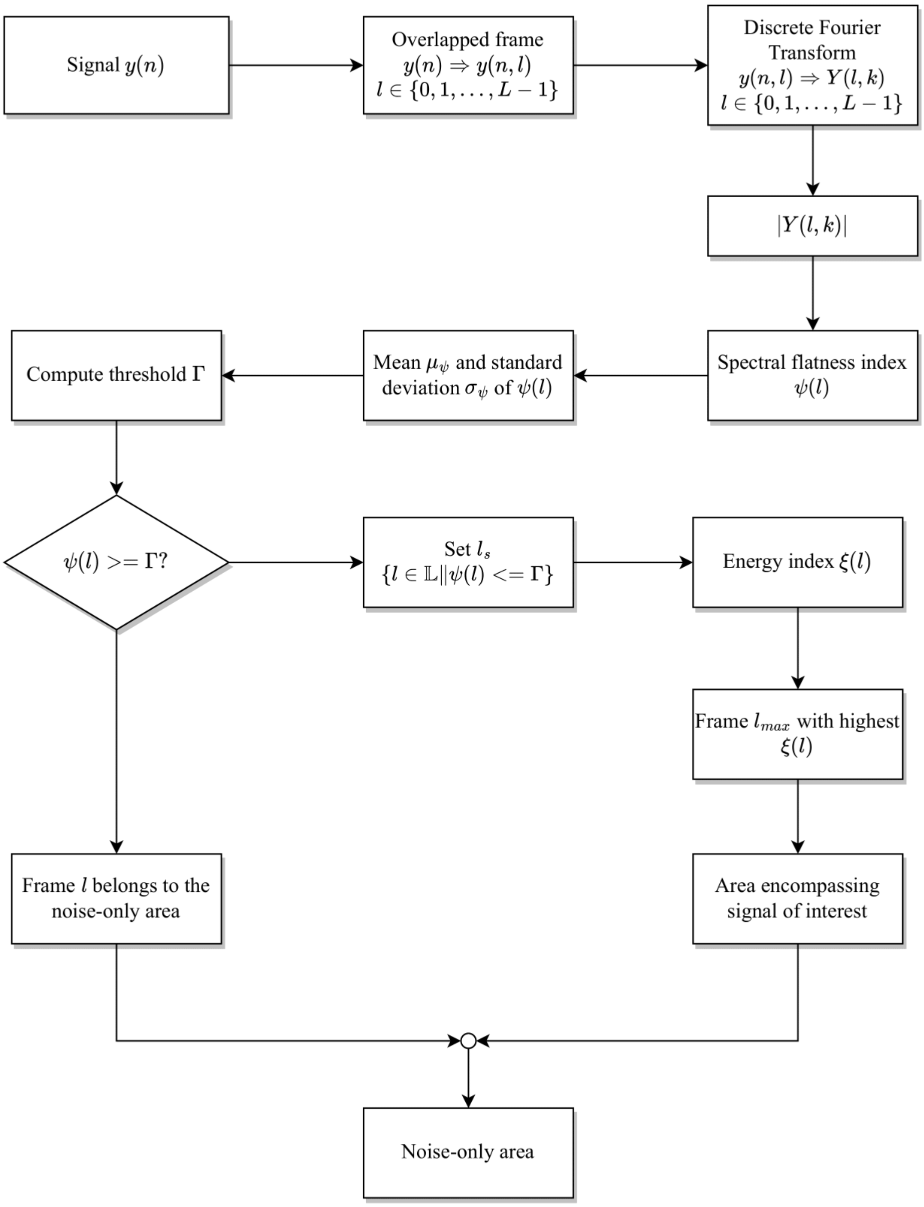
Signal $y(n)$
Overlapped frame $y(n) \Rightarrow y(n,l)$ $l \in \{0,1,\ldots,L-1\}$
Discrete Fourier Transform $y(n,l) \Rightarrow Y(l,k)$ $l \in \{0,1,\ldots,L-1\}$
$|Y(l,k)|$
Spectral flatness index $\psi(l)$
Mean μ_ψ and standard deviation σ_ψ of $\psi(l)$
Compute threshold Γ
$\psi(l) >= \Gamma$?
Set l_s $\{l \in \mathbb{L} || \psi(l) <= \Gamma\}$
Energy index $\xi(l)$
Frame l_{max} with highest $\xi(l)$
Frame l belongs to the noise-only area
Area encompassing signal of interest
Noise-only area

Fig. 1. Block diagram of proposed noise-only area detection.

Assuming that the noise signal, $w(n)$, is an additive noise with which the signal of interest, $x(n)$ is uncorrelated. Thus, the recorded signal $y(n)$, at sample n, can be expressed as:

$$y(n) = x(n) + w(n) \tag{1}$$

As shown in Fig. 1, the first step is to convert $y(n)$ into frames $y(n,l)$ by using 23.22ms Hanning window with 25% overlap at 22050 Hz. The recorded signal in (1) can be written as:

$$y(n,l) = x(n,l) + w(n,l) \tag{2}$$

where $l \in \{0,1,2,\ldots,L-1\}$ is the frame index, L is the number of frames in the recording, $n \in \{0,1,2,\ldots,N-1\}$ and N is the number of samples in each frame. Subsequently, each frame is transformed by applying Discrete Fourier Transform (DFT). The recorded signal can be represented as:

$$Y(l,k) = X(l,k) + W(l,k) \tag{3}$$

where $Y(l,k)$, $X(l,k)$ and $W(l,k)$ are the magnitude spectra of k^{th} frequency component in frame l of the recorded signal, the signal of interest and the noise signal respectively, $k \in \{0,1,2,\ldots,511\}$.The next step of our proposed method is noise-only area detection followed by conducting noise reduction algorithm. These two steps are presented in the below sections.

A. *Proposed noise-only area detection*

The block diagram of the proposed noise-only area detection is shown in Fig. 1. We introduced a noise-only area detection algorithm based on spectral flatness (denoted by ψ). Spectral flatness is a concept that has been applied successfully in many problems of acoustic and speech signal processing. With regards to l^{th} frame, the spectral flatness index $\psi(l)$, can be computed with this expression:

$$\psi(l) = \frac{\sqrt[N]{\prod_{k=0}^{N-1} \left|Y(l,k)\right|}}{\frac{1}{N}\sum_{k=0}^{N-1}\left|Y(l,k)\right|} \tag{4}$$

where $k \in \{0,1,2,\ldots,255\}$. The numerator and the denominator are the geometric and arithmetic mean of 256-point single-sided recorded signal magnitude spectrum respectively.

This index is based on the inequality of the geometric and arithmetic mean as well as the uniformity of the magnitude spectrum obtained by using the short-time Fourier transform. According to [2], the observed signal spectrum appears to exhibit more distinct structure when the desired signal is present, in contrast to when it is absent and this increase in the structure of the signal can be identified by a decrease in the uniformity of the magnitude spectrum in the short-time Fourier representation [2]. In other words, in the frames containing the signal of interest, with different k^{th} component, $Y(k,l)$ exhibits significant variations in value. However, the magnitude spectrum of the short-time Fourier representation for the noise-only frames tends to be flat, indicating that the values are approximately equal. Therefore, $\psi(l)$ of the frames containing only noise tends to be significantly higher compared to time frames with the desired signal.

To use $\psi(l)$ as a criterion to detect noise-only area, we propose a threshold technique. In order to compute this threshold Γ, first we calculate the mean μ_ψ and the standard deviation σ_ψ of the spectral flatness index $\psi(l)$ as below:

$$\mu_\psi = \frac{1}{L}\sum_{l=0}^{L-1}\psi(l) \tag{5}$$

$$\sigma_\psi = \sqrt{\frac{1}{L}\sum_{l=0}^{L-1}\left(\psi(l)-\mu_\psi\right)^2} \tag{6}$$

Assuming that the whole recording is t_1 seconds ($22050t_1$ points at 22050 Hz) long and the signal of interest lasts t_2 seconds ($22050t_2$ points at 22050 Hz). We also assume that $\psi(l)$ follows Normal distribution $\mathrm{N}\left(\mu,\sigma^2\right)$. Thus, the probability that a point belongs to the signal of interest is $\frac{22050t_2}{22050t_1}$ or alternatively $\frac{t_2}{t_1}$. Since one frame has 512 points, the probability that at least M points of a frame belongs to the signal of interest can be written as follows:

$$\mathrm{P}\left(m\right) = \sum_{m=M}^{512}\frac{P_M^{t_2}P_{512}^{t_1-t_2}}{P_{512}^{t_1}} \tag{7}$$

where $P_k^n = \frac{n!}{(n-k)!}$.

Therefore, the probability that with every l so that $\psi(l) <= \Gamma$, at least M points of frame l belongs to the signal of interest can be deduced as follows:

$$\mathrm{P}\left[\psi(l) <= \Gamma\right] = \mathrm{P}\left(m\right) \tag{8}$$

Let $\Phi\left(\Gamma\right) = \frac{1}{\sqrt{2\pi}}\int_\infty^\Gamma e^{t^2/2}\,dt$ gives:

$$\Phi\left(\Gamma\right) = \mathrm{P}\left(m\right) \tag{9}$$

Thus:

$$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{\Gamma}{\sqrt{2}}\right)\right] = \mathrm{P}\left(m\right) \tag{10}$$

Let $\Gamma = \mu_\psi + c\sigma_\psi$ gives:

$$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{\mu_\psi + c\sigma_\psi}{\sqrt{2}}\right)\right] = \mathrm{P}\left(m\right) \tag{11}$$

Therefore, $c = \frac{\sqrt{2}\operatorname{erf}^{-1}\left[2\mathrm{P}\left(m\right)-1\right]-\mu_\psi}{\sigma_\psi}$.

Hence, we have the threshold $\Gamma = \mu_\psi + c\sigma_\psi$ to determine the probable frames whose M points belong to the signal of interest.

Then, we define set l_s as $l_s = \{l \in \mathrm{L} \,||\, \psi(l) <= \Gamma\}$ where L is the set of all frames. For each $l \in l_s$, we compute energy index $\xi(l)$ as follows:

$$\xi(l) = \sqrt{\sum_{k=0}^{L-1}\left|Y(k,l)\right|^2} \tag{12}$$

Let $l_{max} = \underset{l \in \mathrm{L}}{\arg\max}\,\xi(l)$. The frame l_{max} is the frame that contains the peak of the signal of interest. The area encompassing the signal of interest is the set of frames F_S which can be expressed as:

$$\begin{aligned} F_S &= \left\{l \in l_s \mid l <= l_{max}, \left(\forall l' \in \square, l' \in \left[l, l_{max}\right] : l' \in l_s\right)\right\} \\ &\cup \left\{l \in l_s \mid l > l_{max}, \left(\forall l' \in \square, l' \in \left[l_{max}\right] : l' \in l_s\right)\right\} \end{aligned} \tag{13}$$

Therefore, the noise-only area is also able to be represented as a set of frames F_N:

$$F_N = \left\{l \in \mathrm{L} \mid l \notin F_S\right\} \tag{14}$$

B. *Proposed noise reduction algorithm*

Our proposed noise reduction algorithm is based on Noise Spectral Gating, a technique commonly used in sound mixing and manipulation that attenuates a signal according to certain threshold [3]. This method has been employed in the field of general noise removal, with certain modifications. In this selected approach, we evaluate the threshold on the basis of the determined area encompassing the signal of interest and the noise-only area.

First, for every frame of L, we convert their short-time Fourier representation to decibel scale by using this formula: $Y_{dB}(l,k) = 20\log_{10}(Y(l,k))$.

Subsequently, in case of each k^{th} frequency component, we compute the mean $\mu_S(k)$ and the standard deviation $\sigma_S(k)$ of $Y_{dB}(l,k)$ with $l \in F_S$ as follows:

$$\mu_S(k) = \frac{1}{L_S}\sum_{l \in F_S} Y_{dB}(l,k) \qquad (15)$$

$$\sigma_S(k) = \sqrt{\frac{1}{L_S}\sum_{l \in F_S}\left(Y_{dB}(l,k) - \mu_S\right)^2} \qquad (16)$$

where L_S is the number of frames belonging to S.

Similarly, we also compute the mean $\mu_N(k)$ and the standard deviation $\sigma_N(k)$ of $Y_{dB}(l,k)$ with $l \in N$ for each k^{th} frequency component as below:

$$\mu_N(k) = \frac{1}{L_N}\sum_{l \in F_N} Y_{dB}(l,k) \qquad (17)$$

$$\sigma_N(k) = \sqrt{\frac{1}{L_N}\sum_{l \in F_N}\left(Y_{dB}(l,k) - \mu_N\right)^2} \qquad (18)$$

where L_N is the number of frames belonging to N.

To determine whether a frequency component within a frame constitutes a part of the noise signal or the signal of interest, we propose a spectral gate $\Omega(k)$ which can be expressed as:

$$\Omega(k) = \alpha\left(\mu_N(k) + \sigma_N(k)\right) + (1-\alpha)\left(\mu_S(k) - \sigma_S(k)\right) \qquad (19)$$

where $0 <= \alpha <= 1$.

If $Y_{dB}(l,k) <= \Omega(k)$ or $l \in N$, the k_{th} frequency component of the frame l is considered to be a component of the noise signal and $Y_{dB}(l,k) = 0$. Otherwise, $Y_{dB}(l,k)$ remains unchanged.

III. Experiment

A. Data preparation

We visited a local durian orchard in order to collect samples. Our group conducted a tapping experiment with 110 Ri6 durians. Each fruit was tapped ten times per day at the designated spots over two weeks before the ripening date which was defined as the day that durian's husk started to split naturally. Totally, the group collected 4620 original sound recordings. To verify the reliability of our proposed noise reduction algorithm, besides the original recordings, we also conduct experiments on additional samples generated by mixing the original recordings with 6 categories of noise from UrbanSound8K dataset (car horn, children playing, dog bark, engine, drilling and drum sound) [4]. Gunshot sounds from the UrbanSound8K dataset [4] were intentionally introduced to certain additional samples to assess the performance of our algorithm in scenarios where the noise amplitude surpasses that of our target signal. This deliberate inclusion of gunshot sounds allowed us to evaluate how well the algorithm handles situations where the background noise significantly outweighs the desired signal. Fig. 2 illustrates the original sample as well as its variants. These variants were acquired by introducing 5 out of 6 noise sources into the initial recording.

B. Objective measure for noise reduction

Due to the noisy nature of the environment where we performed sample collection, there were a few numbers of audios which were not subject to the interference of background noise. Therefore, we decided to use pattern which was observed in spectral analysis as the ground-truth. The

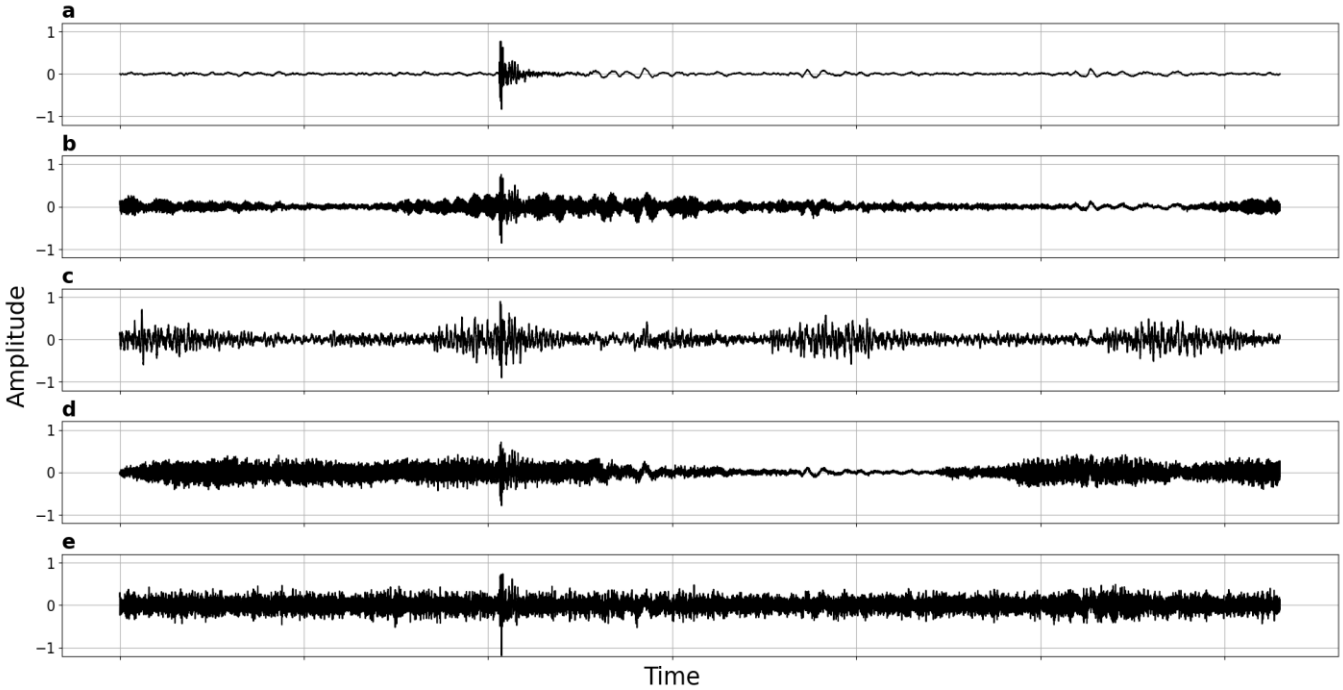

Fig. 2. **a.** Original recording and its variants generated by mixed the original recording with **b.** children playing sound, **c.** drum sound, **d.** car honk sound and **e.** drilling sound.

samples were categorized into 4 groups based on the date of their recording. These 4 groups are as follows:

1. Samples recorded 0 to 3 days before the ripening date.
2. Samples recorded 4 to 6 days before the ripening date.
3. Samples recorded 7 to 9 days before the ripening date.
4. Samples recorded 10 to 14 days before the ripening date.

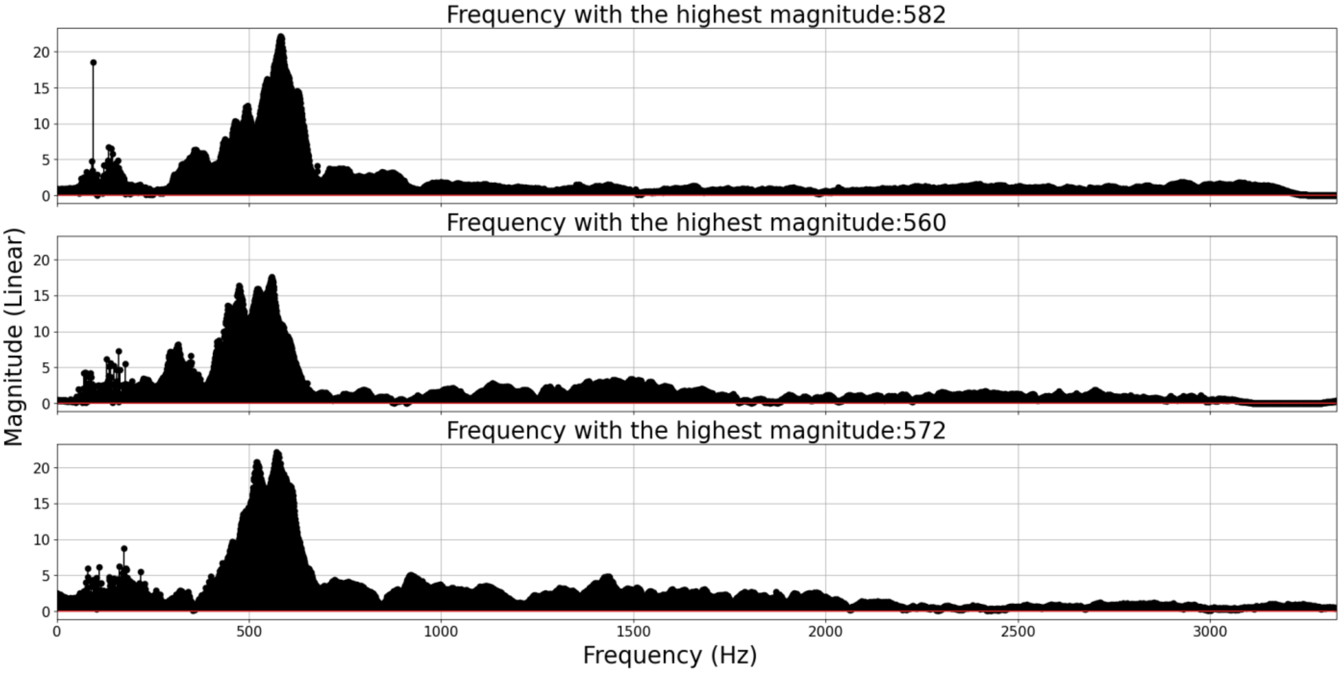

Fig. 3. Single-sided amplitude spectrum of 3 audios recorded at 9 to 14 days before the ripening date.

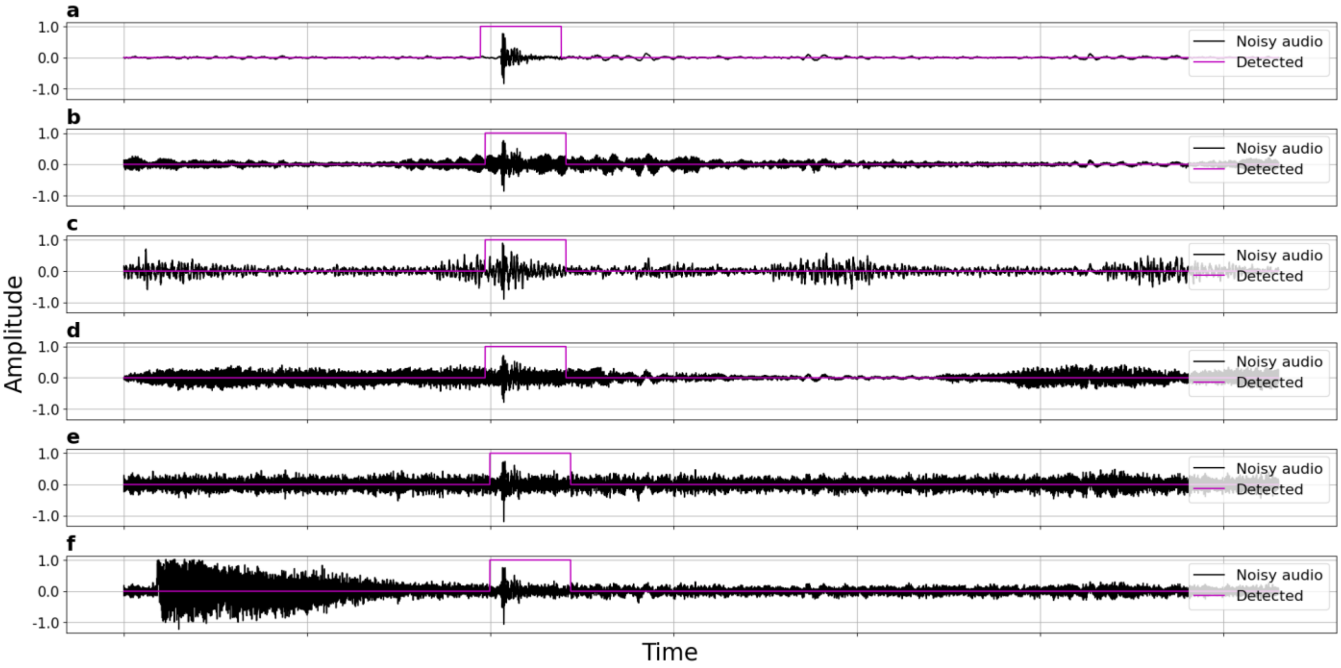

Fig. 4. Noise-only area detection algorithm application on **a.** original recording and its variants generated by mixed the original recording with **b.** children playing sound, **c.** drum sound, **d.** car honk sound, **e.** drilling sound, **f.** gunshot and drum sound (0/1: containing only noise/containing signal of interest).

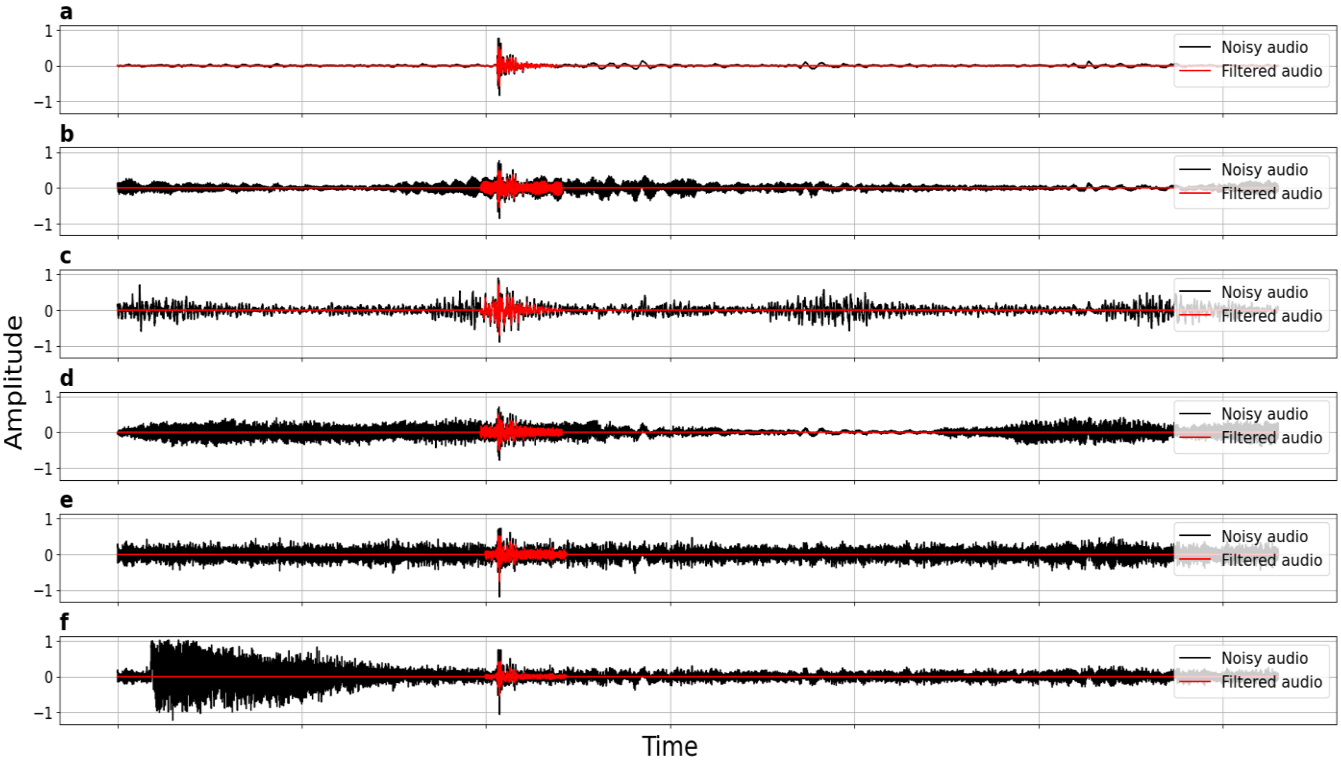

Fig. 5. A comparison between the original noisy audio and the filtered audio reveals the significant reduction of background noise (**a.** original recording and its variants generated by mixed the original recording with **b.** children playing sound, **c.** drum sound, **d.** car honk sound, **e.** drilling sound and **f.** gunshot and drum sound.)

Upon performing Discrete Fourier Transform on these 4 groups, it was observed that recordings (both degraded and not degraded by background noise) of each group all exhibited their respective frequencies with the highest magnitude within a specific range. Three recordings from group 1 were obtained, and these samples were recorded within 9 to 14 days before the ripening date. As shown in Fig. 3, in these recordings, three frequencies with magnitudes at their highest were identified: 582 Hz, 560 Hz, and 572 Hz. These three frequencies were in close proximity to each other in terms of their values. After thoroughly analyzing the original data set, we have identified the range of frequencies with the highest magnitude for each group as Table I.

TABLE I. Range of frequency with highest magnitude.

Group	Range
Samples recorded 0 to 3 days before the ripening date	52-124 Hz
Samples recorded 4 to 6 days before the ripening date	150-267Hz
Samples recorded 7 to 9 days before the ripening date	291-421Hz
Samples recorded 10 to 14 days before the ripening date	531-628Hz

On account of the observation that each group had their own range of value for the frequency with the highest magnitude, we utilized this pattern as our evaluation measure. The proposed noise reduction algorithm's success can be determined by whether a recording's frequency with the highest magnitude falls within its expected range after applying the algorithm. If the frequency is outside the expected range, it indicates that the algorithm has removed the recording's characteristics, resulting in a failure of the noise reduction process. Conversely, if the recording's characteristics are preserved, and the frequency remains within the expected range, it signifies the success of our proposed algorithm. Furthermore, we define that if at least 256 points of a frame belongs to the signal of interest, that frame is also considered to contain the signal of interest itself, which means that in (7), $M = 256$. With regards to the spectral gate $\Omega(k)$ in (19), we arbitrarily choose $\alpha = 0.5$.

IV. Results and Discussion

To illustrate the performance of our noise reduction algorithm, 4620 original recordings and 138600 supplementary recordings (original recordings mixed with different types of noise from UrbanSound4K [4]) are utilized. As illustrated in Fig. 4, despite being affected by noise contamination, our algorithm successfully identifies the regions occupied solely by noise. In particular, in one instance (Fig. 4f), the noise amplitude significantly overshadows that of our target signal, making it challenging to distinguish between the two. Nevertheless, our algorithm accurately estimates and isolates the regions consisting purely of noise.

Furthermore, as demonstrated in Fig. 5, our noise reduction algorithm effectively eliminates the background noise in the noise-only area, even when heavily contaminated (Fig. 5f). Table II showcases the success rate of our proposed algorithm when applied to the original audio as well as audio mixed with various categories of noise. The results demonstrate the robustness of our algorithm across different recording scenarios with various forms of interference. It effectively handles recordings subject to different types of noise, achieving favorable outcomes.

TABLE II. Success rate of proposed noise reduction algorithm

Type of audio	Success	Failure	Success rate
The initial recordings	4620	0	100%
The initial recordings mixed with children sound	23017	83	99.64%
The initial recordings mixed with car horn sound	22775	325	98.59%
The initial recordings mixed with dog bark sound	23078	22	99.9%
The initial recordings mixed with engine sound	22843	257	98.89%
The initial recordings mixed with drilling sound	22754	346	98.5%
The initial recordings mixed with drum sound	23061	39	99.83%

V. Conclusion

In conclusion, this paper introduces a noise reduction algorithm that leverages spectral flatness and spectral noise gate techniques in various noise conditions. The algorithm effectively detects the noise-only regions and the sections containing the desired target signal using spectral-flatness based thresholding. Subsequently, noise gates are computed for each frequency, acting as thresholds to identify whether a frame belongs to the signal or not. Frames identified as containing only noise are then removed.

Experimental results demonstrate the efficacy of the proposed noise reduction algorithm, yielding favorable outcomes and exhibiting reliability for potential future applications. The algorithm's ability to distinguish noise and preserve the target signal makes it a promising approach for addressing noise-related challenges in audio processing and related fields.

Acknowledgment

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting (time and facilities) this study.

References

[1] S. C, C. M L and M. A. Anusuya, “Noise Cancellation and Noise Reduction Techniques: A Review,” in *2019 ICAIT*, Chikmagalur, India, pp. 159-166, 2019.

[2] N. Madhu. “Note on measures for spectral flatness,” *Electronics Letters*, vol. 45, issue 23, pp. 1195–1196, 2009.

[3] J. Hodgson, *Understanding Records: A Field Guide to Recording Practice*, New York: Continuum, pp. 272, 2010.

[4] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in *22nd ACM-MM*, FL, USA, pp. 1041–1044, 2014.