Intelligent Automation And Soft Computing, 2020
Copyright © 2020, TSI® Press
Vol. 26, no. 1, 149–159
https://doi.org/10.31209/2019.100000136

Noise Cancellation Based on Voice Activity Detection Using Spectral Variation for Speech Recognition in Smart Home Devices

Jeong-Sik Park[1] **and Seok-Hoon Kim**[2]

[1]Department of English Linguistics & Language Technology, Hankuk University of Foreign Studies, Seoul, Republic of Korea
[2]Department of Electronic Commerce, Paichai University Studies, Daejeon, Republic of Korea

ABSTRACT
Variety types of smart home devices have a main function of a human-machine interaction by speech recognition. Speech recognition system may be vulnerable to rapidly changing noises in home environments. This study proposes an efficient noise cancellation approach to eliminate the noises directly on the devices in real time. Firstly, we propose an advanced voice activity detection (VAD) technique to efficiently detect speech and non-speech regions on the basis of spectral property of speech signals. The VAD is then employed to enhance the conventional spectral subtraction method by steadily estimating noise signals in non-speech regions. On several experiments, our approach achieved superior performance compared to the conventional noise reduction approaches.

KEY WORDS: Noise cancellation, Non-stationary noise reduction, Smart home device, Spectral subtraction, Speech recognition, Voice activity detection

1 INTRODUCTION

IN recent years, speech recognition technology has been advanced in accordance with technical breakthrough of deep learning technology. The technical advancement achieved successful commercialization of speech recognition, while various smart devices such as smartphones adopted speech recognition functions. In current days, speech recognition technology permeates electronics in home environments such as smart televisions and smart home assistant speakers (Deng and Gong (2014)).

A lot of experts expect that smart home devices will be further evolved along with technical advancement of Internet of Things (IoT) and Artificial Intelligence (AI) technologies. Their progress in functions and services will provide people with more convenient home life by automating living systems in home. And the speech recognition is certainly a core technology of smart home devices by enabling automatic home control and even conveying a communication way between a human and the devices.

To provide more reliable human-machine interaction services based on speech recognition, several problems that negatively affect the recognition performance should be mitigated. The most serious problem is noises surrounded by home environments. The smart home devices are exposed to various types of background noises including music, audio sounds in a television, or talks (Plapous, et. al. (2006), Shrawankar and Thakare (2011)). The noises may significantly degrade the speech recognition performance.

For the past decades, many researchers have attempted to eliminate noise signals contaminating speech signals for the purpose of enhancing the recognition accuracy. A lot of noise cancellation techniques have been introduced, including spectral subtraction (Lu and Loizou (2008)), Wiener filter (Benesty, et. al. (2005), EI-Fattah, et. al. (2014)) Kalman filter (Widrow and Stearns (1985), So and Paliwal (2011)) and minimum mean square error (MMSE) (Ephraim and Malah (1984), Schwerin and Pailwal (2014)). Even though the conventional techniques have been successfully employed for enhancing the speech recognition accuracy, they exhibit different noise cancellation performance according to noise types or characteristics.

This paper is organized as follows. Section 2 introduces several conventional noise cancellation approaches. Section 3 discusses noise cancellation for smart home devices, and Section 4 describes the proposed approach. In Section 5, experimental setup

CONTACT Seok-Hoon Kim ✉ kimshn@pcu.ac.kr

and results are explained. The paper concludes in Section 6.

2 THE CONVENTIONAL NOISE CANCELLATION TECHNIQUES

A lot of studies on reducing variety types of noises including background noise and channel noise have been introduced with the aim of enhancing quality of speech. Several techniques reported successful application to Automatic Speech Recognition (ASR) systems as a pre-processing procedure in which they play a role in recovering the contaminated speech signals.

According to the approaches, the noise signals are usually assumed to be stationary, additive and uncorrelated to speech. The common procedures consist of the estimate of the noise components and the elimination of them in noisy speech. The approaches work well based on the assumption that if the noise signals are stationary, it is possible to estimate noise components in non-speech regions. Therefore, correct detection of non-speech regions is prior to noise cancellation. The Voice Activity Detection (VAD) techniques are generally used for this work. The VAD gives information of voice activity in each frame, thus allowing to estimate noise components, using the following equation.

$$\left|N(\omega,n)\right|^2 = \lambda\left|N(\omega,n-1)\right|^2 + (1-\lambda)\left|X(\omega,n)\right|^2 \tag{1}$$

where $X(\omega,n)$ and $N(\omega,n)$ are the spectrum of the noisy speech and noise signals, respectively. n is the index of the current frame and λ refers to a smoothing coefficient. On consecutive noise frames, this equation becomes true and the VAD indicates a value of zero.

Several studies introduced more sophisticated noise cancellation methods. The most representative method is the spectral subtraction (Boll (1979), Bittu (2016), Martin (1994)). This technique assumes that the noise signals and speech signals are not correlated each other and additively combined in the time domain. Thus, the power spectrum of noisy speech signals can be described as the sum of the noise spectrum and the speech spectrum. The spectral subtraction method considers that if the characteristics of noise signals change slowly compared to speech signals, the noise components estimated on non-speech frames are used to eliminate noise signals contained in speech regions.

Let $x(t)$, $s(t)$ and $n(t)$ be the noisy speech signal, original clean speech signal, and additive noise signal, respectively. According to spectral subtraction approaches, a spectrum of clean speech signals ($\left|\hat{S}(\omega)\right|$) can be estimated by subtracting an average spectrum of noise signals ($\left|\hat{N}(\omega)\right|$) from a spectrum of noisy speech ($\left|X(\omega)\right|$) (Boll (1979)), as follows.

$$\left|\hat{S}(\omega)\right| = \begin{cases} \left|X(\omega)\right| - \left|\hat{N}(\omega)\right|, & \text{if } \left|X(\omega)\right| > \left|\hat{N}(\omega)\right| \\ 0, & otherwise \end{cases} \tag{2}$$

Another enhancement scheme is called Wiener filter (Scalart and Filho (1996)). The Wiener filter also assumes the stationarity characteristics of noise signals to estimate the power spectrum of the clean speech signals and the noise signals. The basic principle of the Wiener filter is to minimize the following expectation value.

$$E((s(n) - \sum_{k=-\infty}^{\infty} \alpha_k \cdot x(n-k))^2) \tag{3}$$

where $s(n)$, $x(n)$, and α_k indicate the clean speech signals, the noisy speech signals, and the filter coefficient, respectively. To realize this filter in the frequency domain, the speech signals and the noise signals are assumed to be normally distributed and uncorrelated each other. This assumption leads the following equations.

$$E(\left|S(\omega)\right|)^2 = E(\left|X(\omega)\right|)^2 - E(\left|N(\omega)\right|)^2 \tag{4}$$

If the expected values disappear, the frequency response of the filter is obtained as follows.

$$\frac{\left|S(\omega)\right|^2}{\left|X(\omega)\right|^2} = \frac{\left|X(\omega)\right|^2 - \left|N(\omega)\right|^2}{\left|X(\omega)\right|^2} \tag{5}$$

Next, the right part of this equation can be generalized as follows.

$$\begin{aligned} H(\omega) &= \frac{\left|X(\omega)\right|^2 - \left|N(\omega)\right|^2}{\left|X(\omega)\right|^2} \\ &= \frac{\left|S(\omega)\right|^2}{\left|N(\omega)\right|^2 + \left|S(\omega)\right|^2} \\ &= \frac{\xi(\omega)}{1+\xi(\omega)} \end{aligned} \tag{6}$$

$\xi(\omega)$ refers to a priori SNR (Signal-to-Noise Ratio) described as follows.

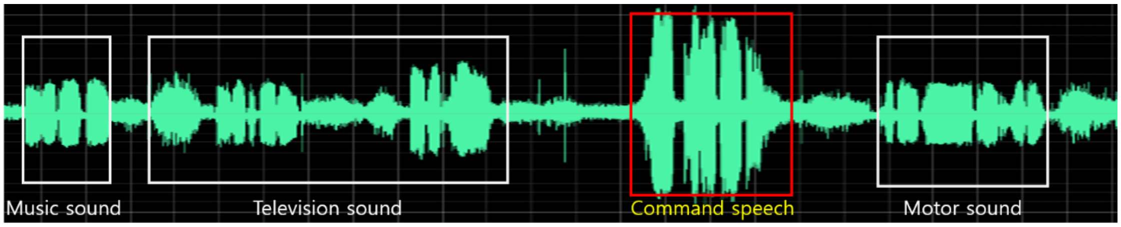

Figure 1. **Noise Signals in Home Environments.**

$$\xi(\omega)=\frac{|S(\omega)|^2}{|N(\omega)|^2} \tag{7}$$

To evaluate the Wiener filter in the spectral domain based on SNRs, this ratio is estimated separately in each frequency. The general posteriori SNR is calculated as:

$$\gamma(\omega)=\frac{|X(\omega)|^2}{|N(\omega)|^2}\,. \tag{8}$$

A priori SNR can be obtained by filtering the noisy speech.

$$\xi(\omega)=\frac{|H(\omega)X(\omega)|^2}{|N(\omega)|^2}=|H(\omega)|^2\,\gamma(\omega) \tag{9}$$

From this equation, both a posteriori SNR and a priori SNR are used to obtain the frequency response of the Wiener filter.

The Mean Squared Error (MSE) has been successfully investigated for speech enhancement. And the Minimum Mean Squared Error (MMSE) estimator minimizing the MSE gave superior performance of noise reduction in log spectral domain (Ephraim and Malah (1984), Malah, et. al. (1999), Kim and Rose (2003)). The MMSE estimator of clean speech signals is described as the power spectrum of noisy speech signals multiplied by a spectral gain function. The function is dependent on frequency and obtained from noise spectrum, SNR, and speech absence probability.

3 NOISE CANCELLATION FOR SMART HOME DEVICES

NOISE signals in home environments have property of non-stationary noise that rapidly and continuously changes, as shown in Figure 1. Types of noises are various, including music, audio sounds in a television, motor sounds of electronic devices, or talks. In addition, for rapid noise cancellation, smart home devices are necessarily required to process noise reduction directly instead of depending on remote server. For the direct processing on devices in real time, computationally low intensive noise cancellation algorithms are suitable for smart home devices. For this reason, we concentrate on spectral subtraction among others, as this technique is known to provide stable performance in spite of relatively low computation.

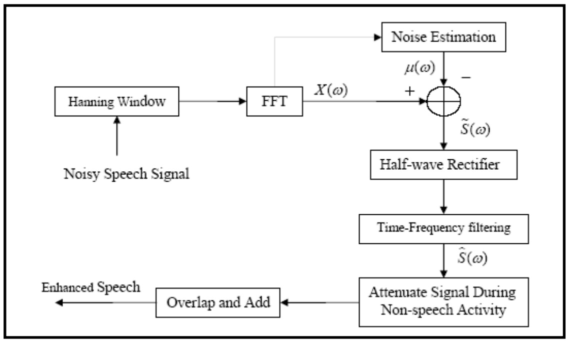

Figure 2. **General procedure of the conventional spectral subtraction.**

3.1 The conventional spectral subtraction

As addressed in Section 2, the spectral subtraction technique assumes that the noise signals and speech signals are not correlated each other and additively combined in the time domain. As the power spectrum of noisy speech signals are to be the sum of the noise spectrum and the speech spectrum, the power spectrum of speech signals is derived by suppressing the noise components estimated on non-speech frames from the power spectrum of noisy speech signals.

$$x(t)=s(t)+n(t) \tag{10}$$

Let $x(t)$, $s(t)$ and $n(t)$ be the noisy speech signal, original clean speech signal, and additive noise signal, respectively.

$$X(\omega)=\hat{S}(\omega)+\hat{N}(\omega) \tag{11}$$

The noisy speech spectrum ($X(\omega)$) is regarded as the sum of the noise spectrum and the speech spectrum. In spectral subtraction methods, the clean speech spectrum ($|\hat{S}(\omega)|$) can be estimated by subtracting the average noise spectrum ($\left|\hat{N}(\omega)\right|$) from the noisy speech spectrum ($\left|X(\omega)\right|$) (Boll (1979)), as mentioned with (2).

Figure 2 shows the general procedure of the conventional spectral subtraction approach. To reduce the distortion of the noisy speech signals, each frame is covered by the Hamming window function. Most of the frames are also calculated by performing the Fast Fourier Transform (FFT) to estimate the spectral component. The first consistent frames assume that noise regions are in signals, and they are then subtracted from every frame of the contaminated signals after estimating the average noise spectrum using their sum. After spectral subtraction, half-wave rectification is performed to remove negative spectral components. When signals are obtained in the spectral domain through such a series of processes, an enhanced signal in the time domain by applying the Inverse Fourier Transform (IFFT).

The conventional spectral subtraction method estimates the average noise spectral energy from consecutive frames of starting point of input signals, assuming that the starting signals are pertinent to non-speech regions. The estimated spectral energy is subtracted for entire signal regions. As shown in this figure, the starting point of input signals (the red section) is assumed to be non-speech regions.

3.2 Drawbacks of spectral subtraction

The conventional method is vulnerable to non-stationary noise although it is quite effective for stationary noise. Therefore, there are two issues.

First, residual noise remains in non-speech regions. Typical spectral subtraction depends on the magnitude of each frame in the spectral domain. The average noise spectrum calculated by the number of first consistent frames is fixed. When it is subtracted sequentially from the entire frame, the magnitude of the current frame may be relatively larger than that of the previous frame. That is, it is necessary to update the spectral components to solve this problem in the non-speech regions.

The second problem is that because of the above problem. Because noise signals and speech signals are not correlated in the contaminated signal, the noise signal magnitudes that are added to the non-speech and speech regions of the clean speech signal are not the same. For this reason, after spectral subtraction, noise reduction in the speech region may not be performed smoothly owing to the negative spectral components that are removed using half-wave rectification.

4 SPECTRAL SUBTRACTION USING SPECTRAL VARIATION BASED VOICE ACTIVITY DETECTION

A main drawback of the conventional spectral subtraction is that dependency on the firstly estimated noise components in a non-speech region may lead to inaccurate subtraction in following speech regions, especially when the noise properties rapidly change. The proposed approach aims to consecutively detect non-speech regions expected to preserve noise signals, estimate spectral components of the noise signals in the region, and then subtract the estimated components from following speech regions with spectral subtraction technique.

In consideration of continuously estimating and updating noise components to be used for spectral subtraction, we attempt to advance the conventional voice activity detection technique for the purpose of more sophisticated detection of non-speech regions for non-stationary noises.

4.1 Voice activity detection based on spectral variation

Various voice activity detection approaches have been introduced for the correct detection of speech regions (Yiming and Rui (2015), Yang, et. al. (2010), Ramirez, et. al. (2007), Moattar and Homayounpour (2009)). The most representative VAD techniques include energy-based approach and zero-crossing rate approach. The energy-based approach considers a general tendency that speech signals preserve higher signal energy than non-speech signals. Based on this property, the approach classifies speech regions that indicate higher energy compared to a pre-determined threshold. The zero-crossing rate means a frequency of changing in between a positive value and a negative value of signals. This approach also considers a general property of signals that higher zero-crossing rate occurs in speech regions compared to non-speech regions. The rate is used to determine speech or non-speech regions.

Even though the conventional VAD techniques have been successfully applied for speech recognition, most of them demonstrated different performance according to input noise signals, in particular, non-stationary noises. The main reason of the performance deterioration is that most of the conventional techniques highly depend on a pre-determined threshold that determines if a given frame is a speech or non-speech region. The most representative example is that energy-based VAD may incorrectly categorize a non-speech region as a speech region, as the threshold adjusted to noiseless signals determines non-speech regions as speech regions, in which background noise signals increase the energy of the non-speech regions more highly than the threshold.

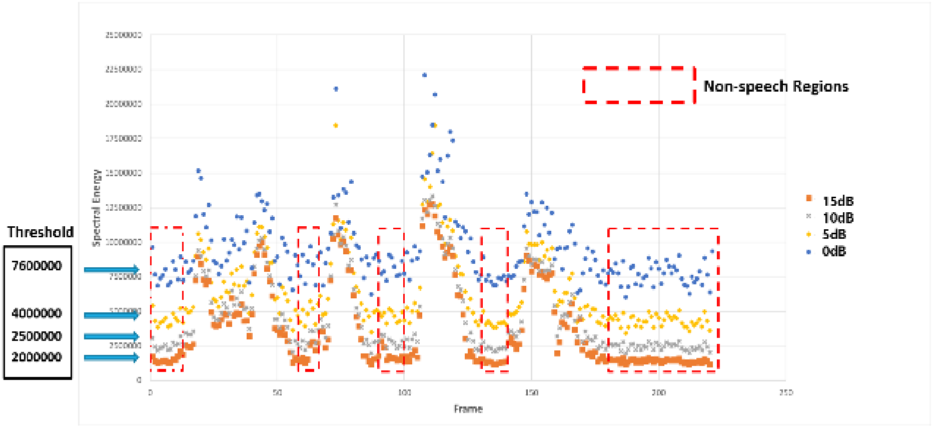

Figure 3. **Optimal Threshold Variation according to Different SNRs (dB) in the Conventional Spectral Energy based VAD.**

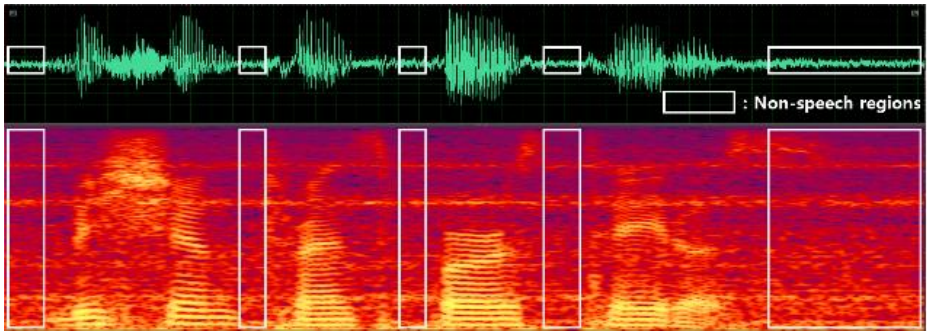

Figure 4. **Spectral Characteristics of Speech and Non-speech Regions of Noisy Speech.**

To solve the drawback of the conventional VAD techniques, the threshold should be changed according to noise levels called signal-to-noise ratio. We examined optimal threshold for VAD to investigate this drawback, by using spectral energy based VAD technique. Figure 3 demonstrates that optimal thresholds estimated for VAD significantly vary according to different SNRs. But defining different thresholds for SNRs does not provide an optimal solution in non-stationary noise environments such as home devices in which SNRs of input noise signals cannot be predicted.

In this study, we propose a way of determining a fixed threshold irrelevant to SNRs. For this work, we investigated spectral characteristics of speech regions and non-speech regions in different SNRs. Figure 4 represents a spectrogram (below) pertinent to a speech waveform (above) recorded in noisy environments. As shown in the spectrogram of this figure, non-speech regions indicate a tendency that spectral energy persists in same frequency bins. On the other hand, in speech regions, the energy varies in each frequency bin according to time.

Concentrating on this spectral tendency, we examined a variation of spectral energy for each of two regions with the standard deviation. First, for each frame of a same length (20ms), spectral energy is calculated in each frequency bin. And then an average of spectral energy values calculated from a certain number of frames (we note a set of the frames as the slice window) is obtained as follows.

$$\mu_l[m] = \frac{1}{L}\sum_{k=0}^{L-1} P_{l,k}[m] \text{ (A)} \qquad (12)$$

where L and l means the number of frames belonging to a slice window and the index of the current slice window, respectively. $P_{l,k}[m]$ is the spectral energy of the m-th frequency bin of the k-th frame in the l-th window. For each frequency bin of the k-th frame, a difference between $P_{l,k}[m]$ and the average is calculated. The difference is then used to calculate the variance over the given slice window, as follows.

$$V_l[m] = \frac{1}{L}\sum_{k=0}^{L-1} (P_{l,k}[m] - \mu_l[m])^2 \qquad (13)$$

Finally, a standard deviation is obtained for each frequency bin. And then the average of the standard deviation values from overall frequency bins is calculated to observe an entire variation of spectral energy for the given slice window, as follows.

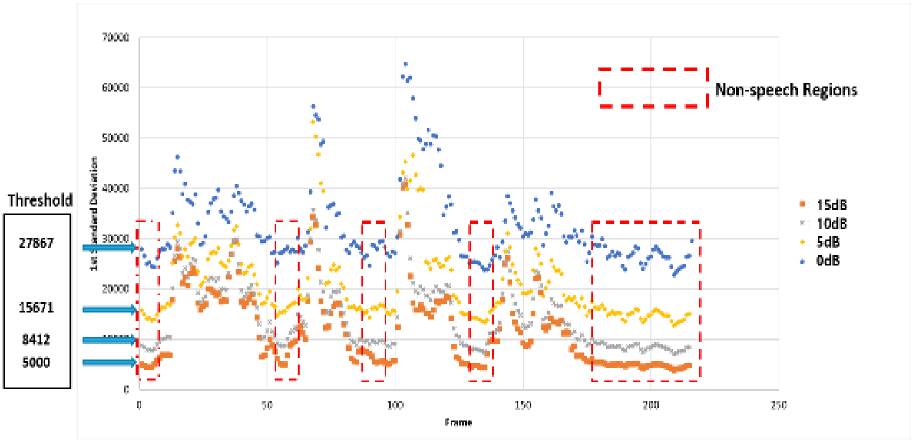

Figure 5. Optimal Threshold Variation according to Different SNRs (dB) in the Standard Deviation based VAD.

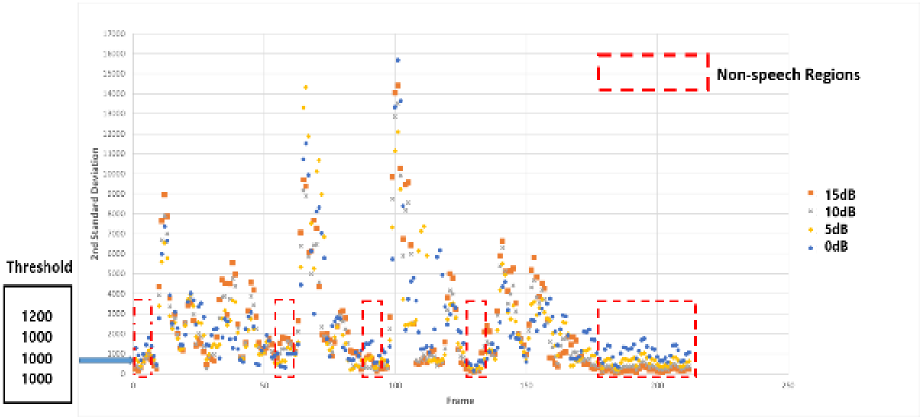

Figure 6. Optimal Threshold Variation according to Different SNRs (dB) in the Second Standard Deviation based VAD.

$$Std_l[m] = \sqrt{V_l[m]} \tag{14}$$

$$Std_l = \frac{1}{N}\sum_{m=1}^{N} Std_l[m] \tag{15}$$

where N means the number of entire frequency bins.

To observe the efficiency of the standard deviation based VAD on different SNRs, we investigated variations of thresholds according to SNRs in the same way as Figure 3. As shown in Figure 5, the standard deviation calculated in non-speech regions was relatively lower than that in speech regions, thus providing sufficient criterion for VAD. In particular, the variation of thresholds on different SNRs was reduced in comparison of the conventional VAD technique based on spectral energy.

Even though the standard deviation represents a very efficient criterion for VAD, the thresholds still vary over SNRs. In order to further reduce variations of thresholds, we examined a variation of the standard deviation for each region, calculating the standard deviation from the first deviation results repeatedly. First, the average of the deviations obtained from each slicing window is calculated, as follows.

$$\mu_l' = \frac{1}{L}\sum_{k=0}^{L-1} Std_{l,k} \tag{16}$$

where $Std_{l,k}$ is the deviation of the k-th frame in the l-th slicing window that was calculated in (15). The prime in this equation denotes the second calculation, distinguishing the average obtained from (12). Next, the variance and the standard deviation for each window are sequentially calculated, as follows.

$$V_l' = \frac{1}{L}\sum_{k=0}^{L-1} (Std_{l,k} - \mu_l')^2 \tag{17}$$

$$Std_l' = \sqrt{V_l'} \tag{18}$$

The efficiency of the secondly estimated standard deviation (shortly denoted as the second deviation) based VAD on different SNRs was validated in the same way as the first deviation-based approach by observing variations of thresholds according to SNRs. As shown in Figure 6, the second deviation calculated in non-speech regions was still relatively lower than that in speech regions, thus satisfying conditions for a criterion for VAD. A remarkable result is that the variation of thresholds on different SNRs was significantly reduced compared to the first deviation-based VAD approach as well as the conventional one.

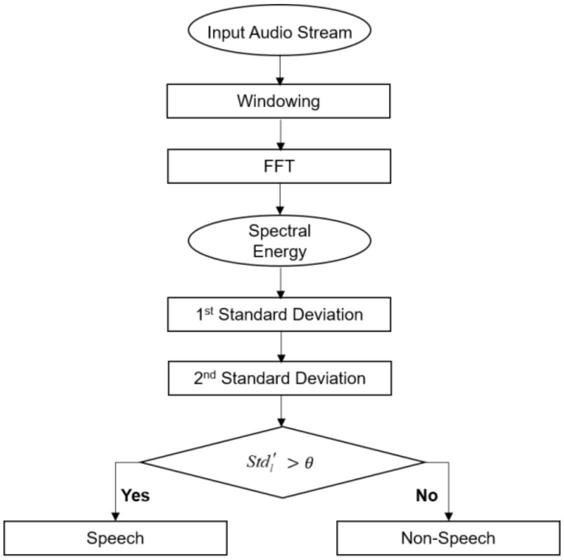

Figure 7. Procedure of the Proposed VAD based on Spectral Variation.

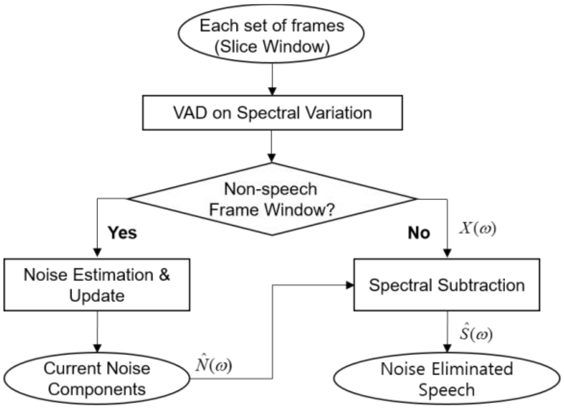

Figure 8. Procedure of the Proposed Noise Cancellation based on Spectral Variation based VAD.

Considering the observation and validation results addressed above, we propose an advanced VAD approach based on variations of spectral energy. Figure 7 represents the procedure of the proposed approach. For every frame, spectral energy of each frequency bin is estimated. The first standard deviation is then calculated from the energy within each slice window. Next, the second deviation is obtained from the first deviation results. It is used for decision of speech regions and non-speech regions by comparing with pre-determined threshold. This threshold is fixed regardless of SNRs.

4.2 *Spectral subtraction for noise cancellation*

The proposed VAD based on spectral variation is used in conducting noise cancellation based on spectral subtraction. Figure 8 illustrates a procedure of the proposed noise cancellation approach. In this figure, $X(\omega)$ denotes the noisy speech signals. $\hat{N}(\omega)$ and $\hat{S}(\omega)$ mean the noise components and de-noised speech signals, respectively.

For a set of frames belonging to each slice window, spectral variation based VAD determines if the region is speech or non-speech. If a given frame window is categorized as a non-speech region, the noise components are estimated and registered as a current noise member. If there is a noise member registered already, the member is updated to current noise frames.

Once a frame window is determined as a speech region, the frames of the window are submitted to a spectral subtraction process. In the noise cancellation process, noise components registered as a current noise member are eliminated from the speech frames according to (2).

5 EXPERIMENTAL SETUP AND RESULTS

5.1 *Experimental setup*

WE performed several experiments to validate the efficiency of the proposed approach. The main purpose of our experiments is to investigate the stability of the noise cancellation performance over SNR variation. For this reason, we used stationary noise data that is suitable for simulating varying SNR conditions. The representative stationary noise data in home environment include air conditioner noise, fan noise, and vacuum noise. Among them, we collected vacuum sounds that convey the severest noise level.

For a fair verification, we compared the performance of the proposed approach with that of the conventional spectral subtraction technique over three varying SNR conditions (15dB, 5dB, 0dB). Naturally, in our approach, the threshold for VAD was fixed as an optimal value obtained using spectral variation based approach. On the other hand, in the conventional technique, optimal threshold values were estimated differently according to three SNR levels.

5.2 Experimental results

There are several ways to investigate noise cancellation performance. A way is to compare the difference between original signals and de-noised signals with visual data such as waveform or spectrogram. Even though the waveform is very intuitive and easy to figure out the difference, it conveys insufficient information over speech regions compared to non-speech regions.

Figure 9 demonstrates the waveform figures of noise-contaminated speech, de-noised speech by conventional spectral subtraction and de-noised speech by the proposed approach. In non-speech regions, a tendency is observed that the conventional and proposed approaches efficiently eliminate noise components of the original signals. In particular, the proposed approach significantly outperforms the conventional subtraction method. However, the performance improvement is not clearly investigated in speech regions.

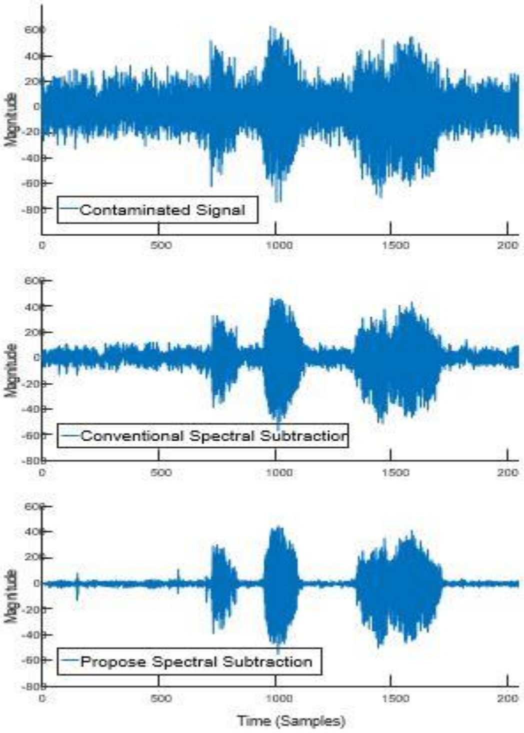

Figure 9. **Noise Cancellation Performance Comparison in Waveform (SNR 5dB).**

Spectrogram figures reflecting spectral energy properties in frequency domain relatively provide more significant performance improvement on both non-speech and speech regions. As shown in Figure 10, the proposed approach remarkably eliminated noise components included in speech regions. Such a tendency was observed in other SNR levels (15dB and 0dB), as shown in Figure 11.

Those visual data proves that the proposed approach successfully outperforms the conventional spectral subtraction technique, although it uses a fixed VAD threshold over different SNR levels. Another method to compare noise cancellation performance is to use particular measures based on mathematical measurement. The representative measures are signal-to-noise ratio and spectral distance (also called spectral distortion).

SNR means a ratio of clean speech signals and the rest of the noise signals after noise cancellation, as follows.

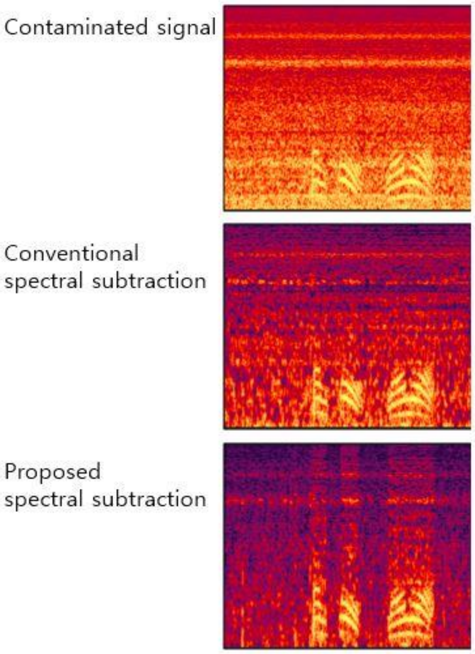

Figure 10. **Noise Cancellation Performance Comparison in Spectrogram (SNR 5dB).**

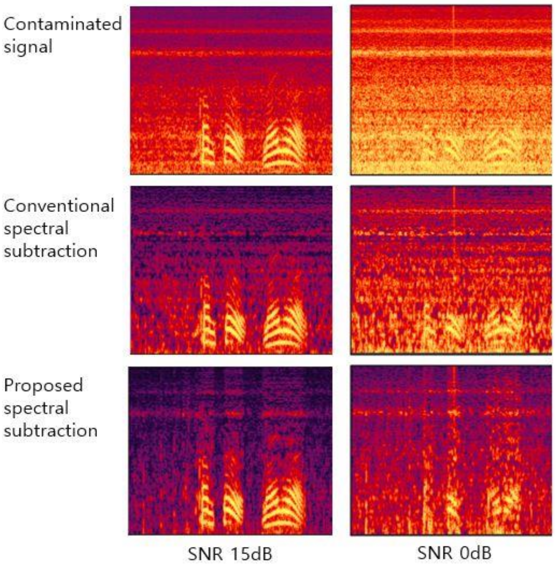

Figure 11. **Noise Cancellation Performance Comparison in Spectrogram (SNR 15dB and 0dB).**

$$SNR = 10\log_{10}\frac{\sum_{t=1}^{L-1} s(t)^2}{\sum_{t=1}^{L-1}(s(t)^2 - \hat{s}(t)^2)} \tag{19}$$

where $s(t)$ and $\hat{s}(t)$ are clean speech signals and de-noised speech signals, respectively. As correctly eliminated the noise signals are, the denominator is closer to 0, increasing the SNR. Figure 12 demonstrates the performance comparison between the proposed and the conventional spectral subtraction. Two approaches improved the SNR levels by successful noise reduction. The proposed approach provided better performance, indicating higher SNRs.

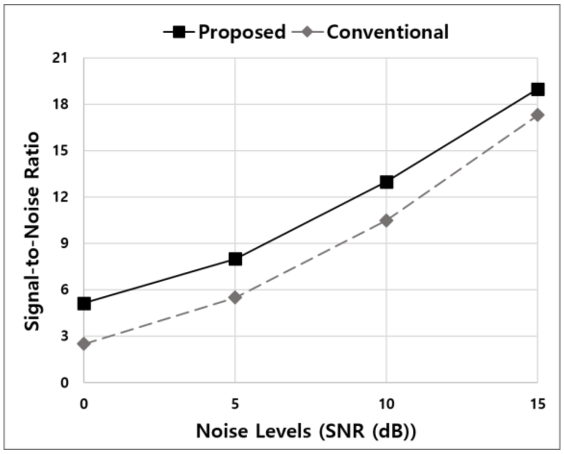

Figure 12. Noise Cancellation Performance Comparison using Signal-to-Noise Ratio Measurement.

The spectral distance (SD) measures the difference between clean speech signals and de-noised speech signals, as follows.

$$SD = \frac{1}{L}\sum_{t=0}^{L-1}\sqrt{\frac{1}{N}\sum_{k=0}^{N-1}\left[10\log_{10}\frac{X(k)^2}{\hat{X}(k)^2}\right]^2} \tag{20}$$

where $X(k)$ and $\hat{X}(k)$ denotes clean speech signals and de-noised speech signals, respectively. N is the number of frequency bins. Hence, this equation calculates the distance using spectral energy in N bins. As perfectly reduced the noise components are, the spectral distance indicates low value. Figure 13 shows the results using spectral distance. As shown in this figure, the proposed approach indicated lower distance over each SNR levels than the conventional method, explaining superior noise cancellation performance.

Two measures certainly proved the efficiency of the proposed approach. A common property shown in two results is that the proposed approach further significantly outperforms on lower SNR levels (0dB and 5dB). This tendency supports very remarkable efficiency of the proposed approach in eliminating severe noise signals.

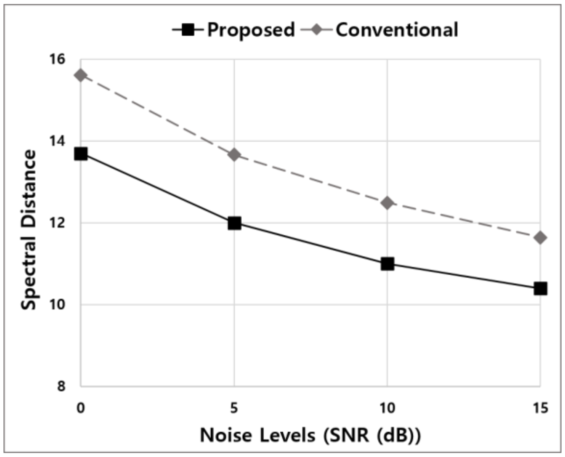

Figure 13. Noise Cancellation Performance Comparison using Spectral Distance Measurement.

6 CONCLUSION

THIS paper proposed an efficient noise cancellation approach for smart home devices. The proposed method aims to enhance the conventional spectral subtraction, using an advanced voice activity detection (VAD). The conventional VAD methods may vulnerable to home environment noises, representing different thresholds according to SNRs. The proposed VAD uses a spectral variation difference between noise signals and speech signals. The advanced VAD steadily detects non-speech regions and estimates noise components. Currently estimated noise components are used for spectral subtraction.

We verified the efficiency of the proposed approach by conducting several experiments. Our approach demonstrated superior performance compared to the conventional approaches.

7 ACKNOWLEDGMENT

THIS work was supported by Hankuk University of Foreign Studies Research Fund, Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(NRF-2017R1D1A1A09000903), the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2016-0-00313) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation), and the research grant of Pai Chai University in 2019.

8 REFERENCES

J. Benesty, J. Chen, Y. A. Huang, and S. Doclo, (2005). Study of the Wiener filter for noise reduction, *Speech Enhancement*. 9-41.

K. Bittu, (2016). Mean-median based noise estimation method using spectral subtraction for speech enhancement technique, *Indian Journal of Science and Technology*. 9(35).

S. F. Boll, (1979). Suppression of acoustic noise in speech using spectral subtraction, *IEEE Transactions on Acoustics, Speech and Signal Processing*. 27(2), 113–120.

M. A. A. EI-Fattah, M. I. Dessouky, A. M. Abbas, A. M. Diab, S. M. EI-Rabaie, W. AI-Nuaimy, S. A. Alshebeili, and F. E. Abd Ei-samie, (2014). Speech enhancement with an adaptive Wiener filter, *International Journal of Speech Technology*. 17(1), 53-64.

Y. Ephraim and D. Malah, (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 32(6), 1109-1121.

H. K. Kim and R. C. Rose, (2003). Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments, *IEEE Transactions on Speech and Audio Processing*. 11(5), 435-446.

J. Li, L. Deng and Y. Gong, (2014). An overview of noise-robust automatic speech recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 22(4), 745-777.

Y. Lu and P. C. Loizou, (2008). A geometric approach to spectral subtraction, *Speech Communication*. 50(6), 453-466.

D. Malah, R. V. Cox, and A. J. Accardi, (1999). Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 201-204.

R. Martin, (1994). Spectral subtraction based on minimum statistics, *Proceedings of IEEE European Signal Processing Conference*. 1182-1185.

M. H. Moattar and M. M. Homayounpour, (2009). A simple but efficient real-time voice activity detection algorithm, *Proceedings of IEEE European Signal Processing Conference*.

C. Plapous, C. Marro, and P. Scalart, (2006). Improved signal-to-noise ratio estimation for speech enhancement, *IEEE Transactions on Audio, Speech and Language Processing*. 14(2), 2098-2108.

J. Ramirez, J. M. Górriz, and J. C. Segura, (2007). Voice activity detection. fundamentals and speech recognition system robustness, *Robust Speech Recognition and Understanding*.

P. Scalart and J. Filho, (1996). Speech enhancement based on a priori signal to noise estimation, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 629-632.

B. Schwerin and K. Pailwal, (2014). Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement, *Speech Communication*. 58, 49-68.

U. Shrawankar and V. M. Thakare, (2011). Adverse conditions and ASR techniques for robust speech user interface, *IJCSI International Journal of Computer Science Issues*. 8(3), 440-449.

S. So and K. K. Paliwal, (2011). Modulation-domain Kalman filtering for single-channel speech enhancement, *Speech Communication*. 53(6), 818-829.

N. Upadhyay and A. Karmakar, (2012). A perceptually motivated multi-band spectral subtraction algorithm for enhancement of degraded speech, *Proceedings of IEEE International Conference on Computer and Communication Technology (ICCCT)*. 340-345.

B. Widrow and S. D. Stearns, (1985). *Adaptive Signal Processing*. Prentice Hall.

X. Yang, B. Tan, J. Ding, J. Zhang, and J. Gong, (2010). Comparative study on voice activity detection algorithm, *Proceedings of IEEE International Conference in Electrical and Control Engineering*. 599-602.

S. Yiming and W. Rui, (2015). Voice activity detection based on the improved dual-threshold method, *Proceedings of IEEE International Conference on Intelligent Transportation in Big Data and Smart City (ICITBS)*. 996-999.

9 DISCLOSURE STATEMENT

NO potential conflict of interest was reported by the authors.

10 NOTES ON CONTRIBUTORS

J. S. Park received his B.E. degree in Computer Science from Ajou University, South Korea in 2001 and his M.E. and Ph.D. degree in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 2003 and 2010, respectively. He is now an associate professor in the Department of English Linguistics & Language Technology, Hankuk University of Foreign Studies. His research interests include speech signal processing, speech recognition, and voice interface for human-computer interaction.

S. H. Kim was an assistant professor in Mobile Media at Suwon Women’s University in 2012 and 2017. He is currently an assistant professor in the Electronic Commerce at Paichai University. His teaching and research specialties are in the fields Mobile computing, Web-App programming, Web-Database, information security.