Journal of Theoretical and Applied Information Technology
28th February 2018. Vol.96. No 4
© 2005 – ongoing JATIT & LLS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# ADAPTIVE SPECTRAL SUBTRACTION FOR ROBUST SPEECH RECOGNITION

[1]JUNG-SEOK YOON, [2]JI-HWAN KIM, [3]JEONG-SIK PARK

[1]Department of Information & Communication Engineering, Yeungnam University, South Korea

[2]Department of Computer Science and Engineering, Sogang University, Seoul Korea

[3]Dep. of English Linguistics & Language Technology, Hankuk University of Foreign Studies, South Korea

E-mail: [1]js920215@ynu.ac.kr, [2]kimjihwan@sogang.ac.kr, [3]parkjs@kaist.ac.kr

## ABSTRACT

Speech recognition rate degrades drastically in extreme noisy environments. Spectral subtraction is one of the representative noise reduction method, but it is vulnerable to non-stationary noise although it is quite effective for stationary noise. In this paper, we propose an adaptive spectral subtraction method to improve the speech recognition performance. The proposed method is to consistently update the noise component in non-speech regions and remove the corresponding component in following speech regions. To validate of the noise reduction performance, we conducted several experiments for each noise power level. Our approach achieved better performance compared to the conventional spectral subtraction approach.

**Keywords:** *Noise Reduction, Spectral Subtraction, Speech Recognition, Voice Activity Detection*

## 1. INTRODUCTION

Recently, there has been an increased tendency to use applications such as personal assistant and smart home devices with speech recognition technology [1]. In particular, they are controlled by a user's voice without directly moving, and are connected to devices based on speech recognition in home environments. However, there is a problem with respect to the significantly degraded speech recognition rate in very noisy environments such as with TVs and when there are conversations between people [2][3]. This may result in serious malfunctions in applications that recognize a user's spoken voice. In this case, it is important to remove noise in the surrounding environment in advance to enable the accurate detection of voice. This improves the speech recognition performance when controlling devices that recognize commands spoken by a user.

Such a technique is called noise reduction, and the removal of noise from contaminated signals is important to achieving advanced pre-processing of voice detection process to improve speech recognition performance. In general, the types of noise are classified as stationary and non-stationary. Stationary noise refers to noise that hardly changes over time, e.g., engine noise in a car. Non-stationary noise refers to noise that fluctuates with time, e.g. engine noise in a fighter plane and conversations between people. Several approaches to noise removal have been extensively studied for a long time. Stationary noise reduction methods include Spectral Subtraction [4], Wiener filter [5][6], minimum mean square error (MMSE) short-time spectral amplitude estimation [7][8], Kalman filter [9][10]. There are also Average Estimator Least Mean Square (AELMS) methods [11], of which the spectral subtraction technology is the most common and representative noise reduction method. However, this approach is vulnerable to non-stationary noise in that noise remains in non-speech regions after subtracting the average noise spectrum from a noisy speech spectrum. In this paper, we propose an adaptive spectral subtraction technique to improve noise reduction performance. The proposed involves consistently updating noise components in non-speech regions.

This paper is organized as follows. Section 2 explains the conventional spectral subtraction method. Section 3 explains the proposed spectral subtraction approach. Experimental results are shown in Section 4, and the originality of the proposed approach is discussed in Section 5. This paper concludes in Section 6.

[*] Corresponding author: Jeong-Sik Park

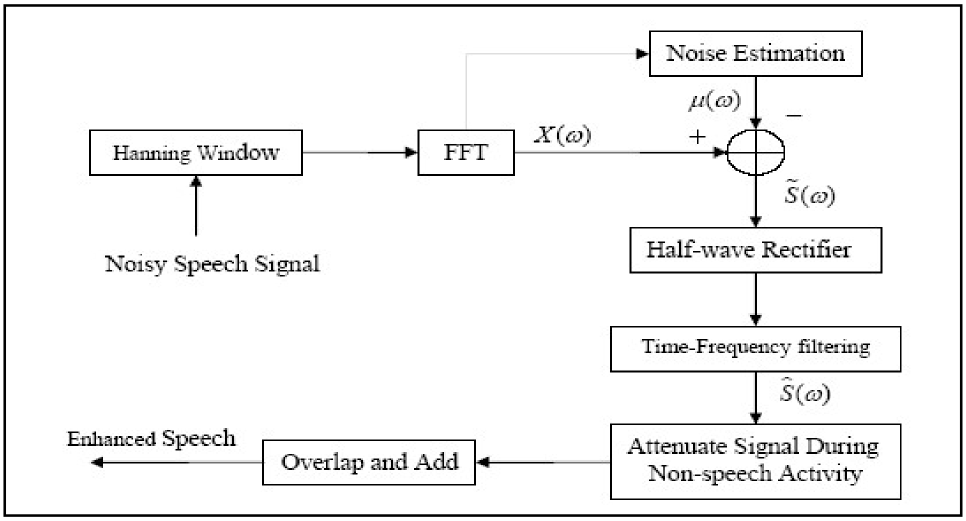

*Figure 1: General procedure of the Conventional Spectral Subtraction*

## 2. THE CONVENTIONAL SPECTRAL SUBTRACTION

### 2.1 Spectral Subtraction

The spectral subtraction technique assumes that the noise and speech are uncorrelated and additive in the time domain [12]. In this case, the power spectrum of the noisy speech signal is the sum of the noise and speech spectra. This method also assumes that noise characteristics change slowly relative to those of speech signals, so that the noise spectrum estimated for non-speech frames can be used to suppress noise components that contaminate the speech.

$$x(t) = s(t) + n(t) \tag{1}$$

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

$$X(\omega) = \hat{S}(\omega) + \widehat{N}(\omega) \tag{2}$$

The power spectrum of the noisy speech signal ($X(\omega)$) is regarded as the sum of the noise and the speech spectra.

$$|\widehat{S}(\omega)| = \begin{cases} |X(\omega)| - |\widehat{N}(\omega)|, & if\ |X(\omega)| - |\widehat{N}(\omega)| \\ 0 \quad , & otherwise \end{cases} \tag{3}$$

In spectral subtraction methods, the clean speech spectrum ($|\hat{S}(\omega)|$) can be estimated by subtracting the average noise spectrum ($|\widehat{N}(\omega)|$) from the noisy speech spectrum ($|X(\omega)|$) [13].

Figure 1 shows the general procedure of the conventional spectral subtraction approach. To reduce the distortion of the noisy speech signal, the Hamming window is applied to each frame. 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).

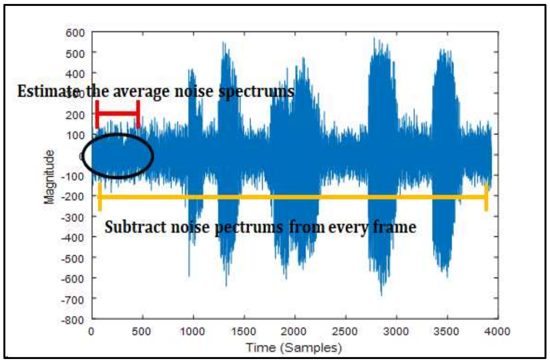

*Figure 2: Principles of the Conventional Spectral Subtraction Method*

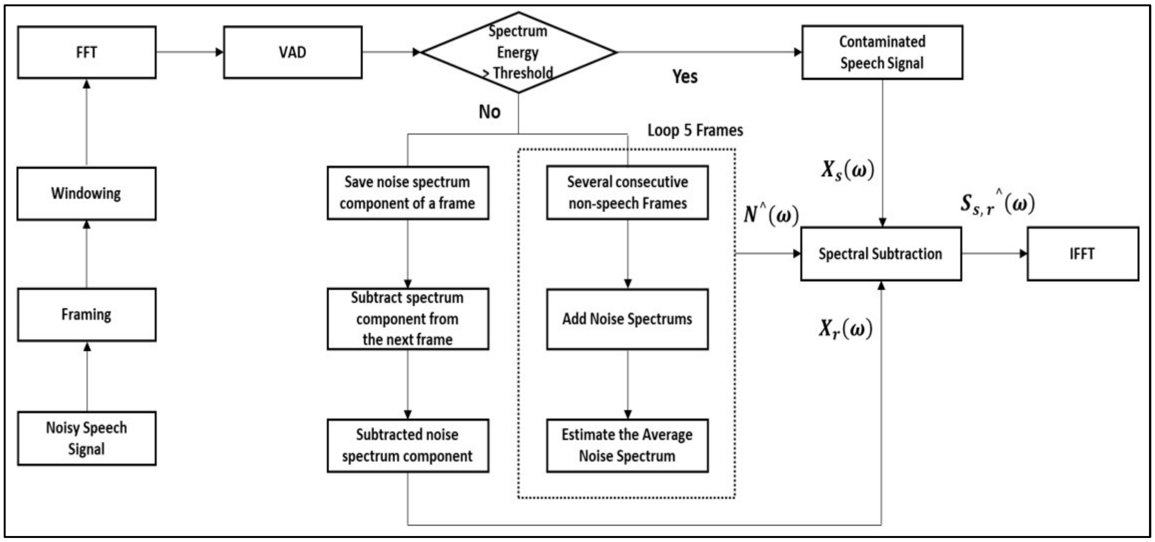
FFT
VAD
Spectrum Energy > Threshold
Contaminated Speech Signal
Yes
No
Loop 5 Frames
Windowing
Save noise spectrum component of a frame
Several consecutive non-speech Frames
$X_s(\omega)$
$N^{\wedge}(\omega)$
$S_{s,r}{}^{\wedge}(\omega)$
Spectral Subtraction
IFFT
Framing
Subtract spectrum component from the next frame
Add Noise Spectrums
$X_r(\omega)$
Noisy Speech Signal
Subtracted noise spectrum component
Estimate the Average Noise Spectrum

*Figure 3: Procedure for the Proposed Spectral Subtraction*

Figure 2 represents the principle of the conventional spectral subtraction method. The 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.

### 2.2 Limitations 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 and speech signals are uncorrelated 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.

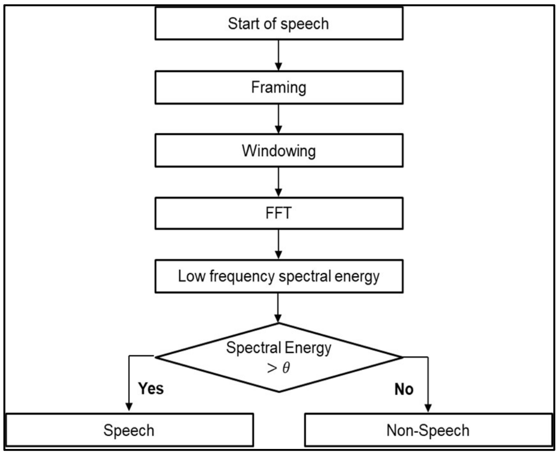

*Figure 4: Procedure for the Voice Activity Detection*

## 3. ADAPTIVE SPECTRAL SUBTRACTION

The spectral subtraction technique is the most common and representative noise reduction method. In this paper, we discussed two issues that affect conventional spectral subtraction. To solve these problems, we propose the adaptive spectral subtraction method, which involves subtracting noise components of several consecutive frames, which are calculated in non-speech regions from speech regions as well as continuously updating it by subtracting them from the non-speech regions.

Figure 3 demonstrates the proposed spectral subtraction procedure. The input noisy speech signals are divided into fixed lengths, and Hamming windows are applied to each frame. Then, they are processed by FFT to obtain spectral components. In addition, as represented in Figure 4, they are classified into speech or non-speech frames using voice activity detection, which compares the energy with a predetermined low frequency energy threshold. In the case of non-speech regions, it performs two functions. In the first case, it subtracts the noise spectrum components of the current frame from the spectral value of the next frame. In the second case, noise components of several consecutive non-speech frames are added, and the average noise spectrum is estimated. Then, spectral subtraction is performed for the contaminated signals and subtracted noise signals, respectively.

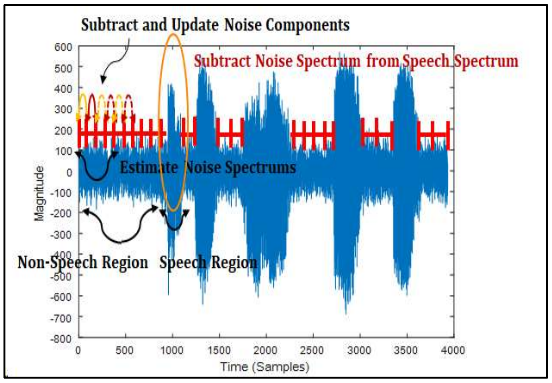

*Figure 5: Principles of the Proposed Method*

Figure 5 represents the principle of the proposed method. The red sections indicate non-speech regions. The proposed method continuously detects non-speech regions and flexibly estimates the noise energy from the regions and the noise components estimated for each non-speech frame are subtracted in the following speech frames.

The decision for the two regions can be expressed as follows.

$$P_t = \sum_{m=0}^{K/d} \left( \sqrt{(Re_t[m]^2 + Im_t[m]^2)} \right)$$
$$P_t > Threshold \ \rightarrow Speech$$
$$P_t \leq Threshold \ \rightarrow Non - Speech \qquad (4)$$

where $P_t$ is total spectral energy based on low frequency bandwidth. $Re_t[m]$ and $Im_t[m]$ are respectively the real and imaginary components of the $m$–$th$ FFT point samples in the $i$–$th$ frame. $t$, $t-1$ denotes the current and previous frames, respectively. $K$ is the total number of FFT points. $d$ means the range of low frequency regions. If the low frequency energy of a frame is higher than the threshold, it is classified as the speech region; otherwise, it is the non-speech region.

As shown in Eq. (5), there is related to the proposed method in this study. If the spectral energy of a current frame for $t$ is smaller than that of the thresholds, the noise component of the previous frame $t-1$ is subtracted from the noise spectral components of the current frame, and can be expressed as follows.

$$\begin{cases} Re_{r,t}[m] = |Re_t[m]| - |Re_{t-1}[m]| \\ Im_{r,t}[m] = |Im_t[m]| - |Im_{t-1}[m]| \end{cases}$$
$$P_t \leq Threshold \ \rightarrow Non - Speech \qquad (5)$$

where $Re_{r,t}[m]$ and $Im_{r,t}[m]$ are the noise spectral components, which are subtracted.

For the recently consecutive non-speech frames in the non-speech region, the average noise spectrum $\widehat{N}(\omega)$, can be obtained as follows.

$$\widehat{N}(\omega) = \frac{1}{L} \sum_{l=0}^{L-1} \left( \sum_{m=0}^{K-1} \sqrt{(Re_t[m]^2 + Im_t[m]^2)} \right) \qquad (6)$$

where $L$ is the total number of fixed frames. $\omega$ is in terms of samples in spectral domain. As represented in Eq. (6), the average noise spectrum is estimated by adding up the spectral energy of recent $L$ frames and dividing by total number of frames in non-speech region.

For the contaminated signals, $X_s(\omega)$ can be obtained as follows,

$$X_s(\omega) = \sum_{m=0}^{K-1} \left( \sqrt{(Re_t[m]^2 + Im_t[m]^2)} \right) \qquad (7)$$

If the spectral energy of the current frame for $t$ is higher than that of the threshold, it is categorized as a speech region.

The calculation of the spectral energy for a noise signal in the current non-speech region can be expressed as follows.

$$X_r(\omega) = \sum_{m=0}^{K-1} \left( \sqrt{(Re_{r,t}[m]^2 + Im_{r,t}[m]^2)} \right) \qquad (8)$$

where $X_r(\omega)$ is the spectral energy calculated with updated components after subtracting the noise spectral components of the previous frame from the current frame.

To estimate the enhanced signals using $X_s(\omega)$, $X_r(\omega)$ and $\widehat{N}(\omega)$, which are obtained from the above equation, $\hat{S}_{s,r}(\omega)$ can be calculated as follows.

$$|\hat{S}_{s,r}(\omega)| = \begin{cases} |X_s(\omega)| - |\widehat{N}(\omega)|, & P_t > \theta \\ |X_r(\omega)| - |\widehat{N}(\omega)|, & P_t \leq \theta \\ 0 & , \mathit{Otherwise} \end{cases} \qquad (9)$$

where $\theta$ is the predetermined threshold. If the spectral energy of the current frame is higher than $\theta$, the noise spectral energy from contaminated signal is subtracted from updated noise signal.

## 4. EXPERIMENTAL SETUP AND RESULTS

In this section, we present experimental results for the performance of the adaptive spectral subtraction approach proposed in Section 3.

### 4.1 Experimental Setup

In this study, we conducted an offline test to evaluate noise reduction performance. The threshold value used in the VAD was empirically estimated to determine the predetermined threshold. An original speech signal is a clean signal without noise. In the experiment, we used four kinds of the noisy speech signals, and noise signals were non-stationary vehicle noise at SNRs of 0, 5, 10, and 15 dB. For each frame, the number of samples was 320, and the frame length was fixed at 20 ms. There were 512 FFT points, and a Hamming window was applied to each frame to prevent signal distortion. The sampling frequency was set to 16 kHz.

### 4.2 Experimental Results

We compared the noise reduction performance using the signal waveforms with the spectrograms, and we investigated to determine the extent of the removal of noise between the conventional method and proposed method. To do this, we tested for each noise power level. We also used the low frequency energy based VAD to determine the speech and non-speech regions for the input noisy speech signal in this study.

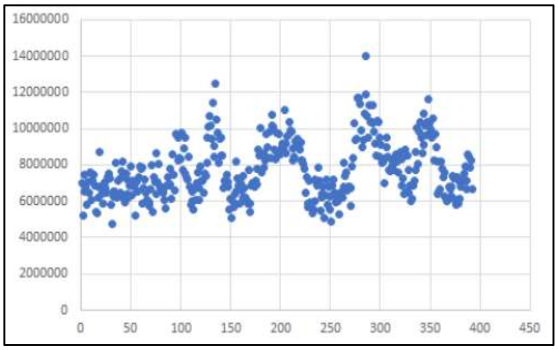

*Figure 6: Distribution of low frequency energy at 0 dB*

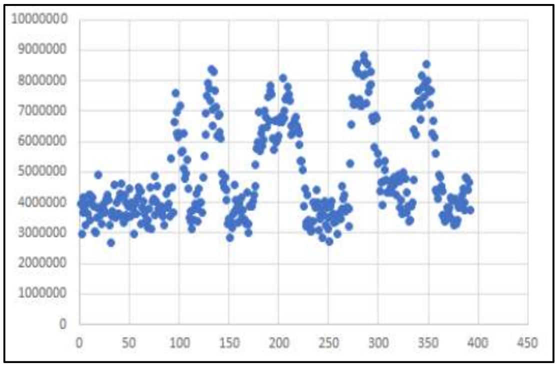

*Figure 7: Distribution of low frequency energy at 5 dB*

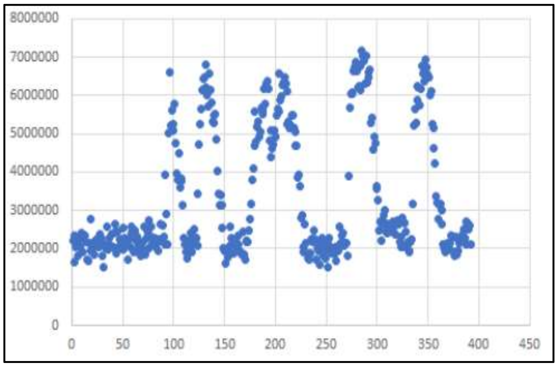

*Figure 8: Distribution of low frequency energy at 10 dB*

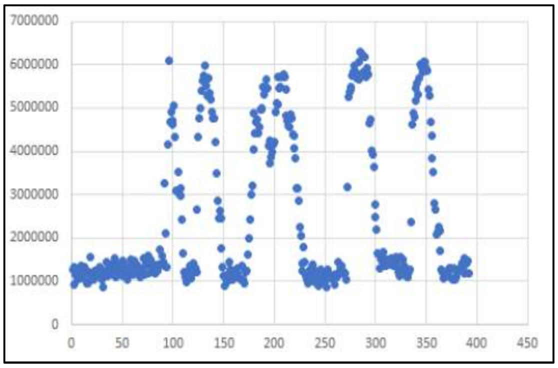

*Figure 9: Distribution of low frequency energy at 15 dB*

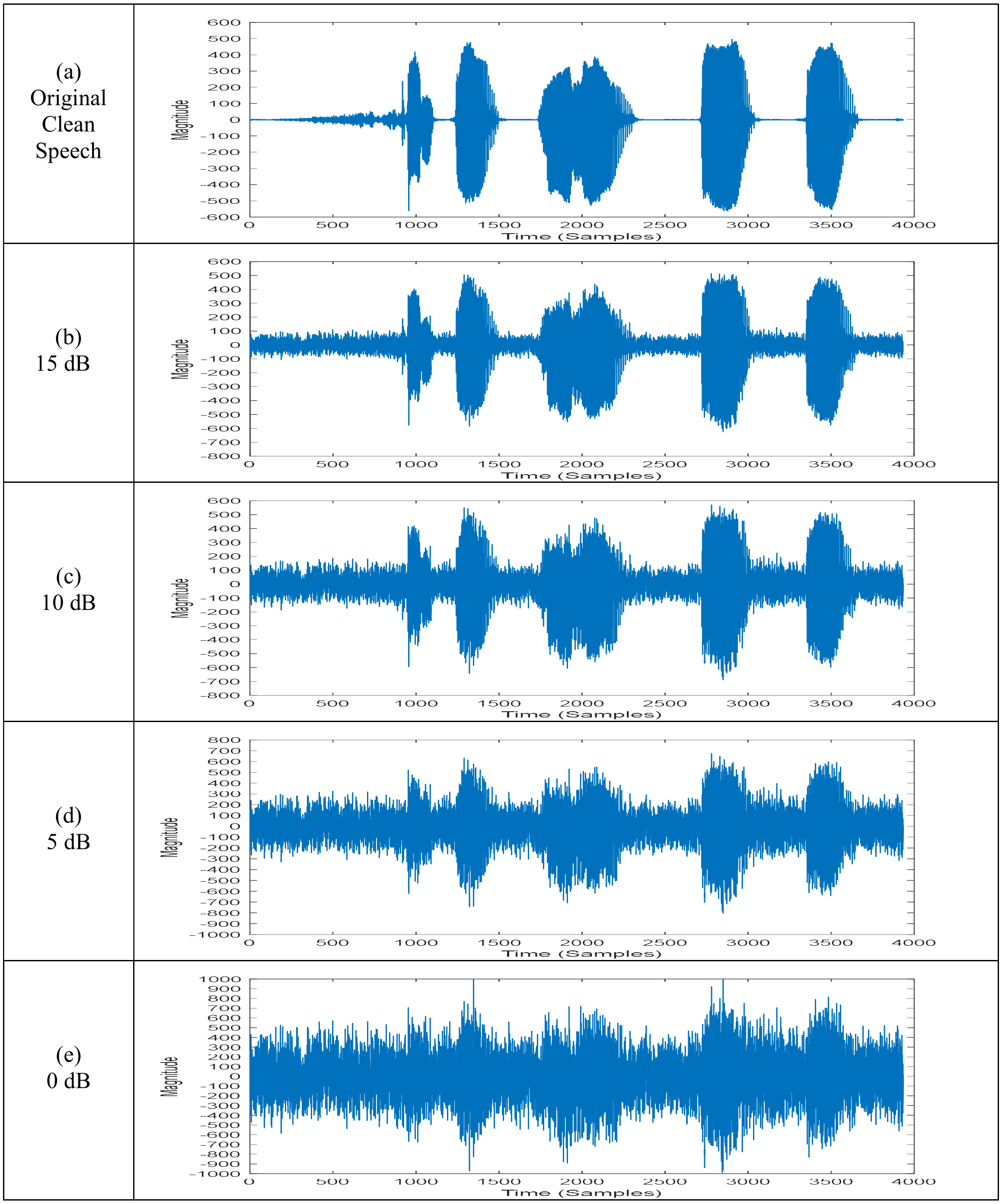

*Figure 10: Speech signal for words, "Come to manual three two," which are spoken by a male speaker: The above pictures represent the noisy speech signals that were added for each noise power level. (a) Original Clean Speech; (b) Noisy Speech at 15 dB; (c) Noisy Speech at 10 dB; (d) Noisy Speech at 5 dB; (e) Noisy Speech at 0 dB*

Figures 6, 7, 8, and 9 show the distribution of the low frequency energy according to each value of noise power level. As shown in the figures, the lower the noise power, the greater is the magnitude of the spectral energy, and the smaller is the distance between speech region and non-speech region. In particular, as shown in the distribution at 0 dB, the distance is very small compared to the others. This is not only difficult to find a threshold for determining two regions, but also result in the loss of the speech signal, which is subtracted from the noise energy by being categorized as a noise region.

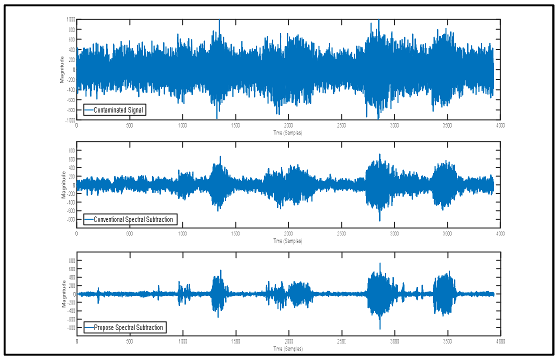

*Figure 11: Results of processing Noisy Speech (0dB)*

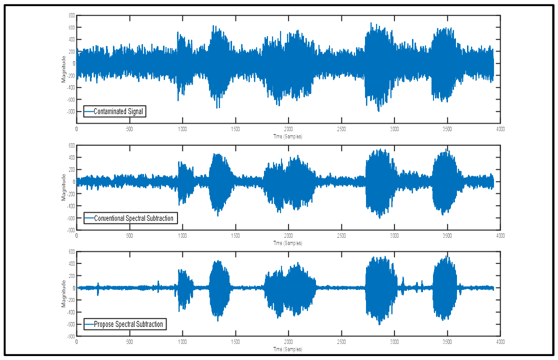

*Figure 12: Results of processing Noisy Speech (5dB)*

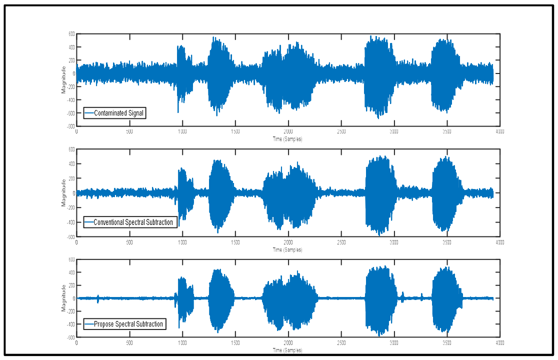

*Figure 13: Results of processing Noisy Speech (10dB)*

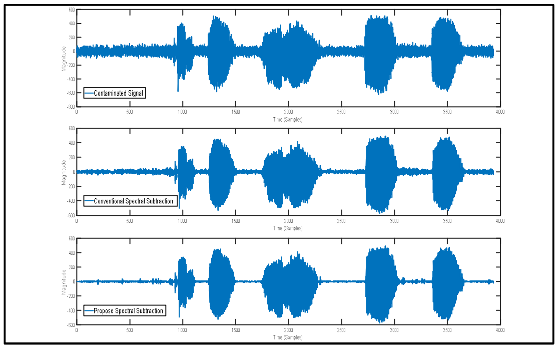

*Figure 14: Results of processing Noisy Speech (15dB)*

We investigated waveforms of noisy speech and de-noised speech signals for performance comparison. Figure 10 illustrates original clean speech and noise-contaminated speech signals. Figures 11, 12, 13, and 14 also represent signal results obtained for the conventional method and proposed adaptive spectral subtraction approach for contaminated signals. As shown in the figures, the first picture is the noisy speech signal for each noise power levels, and the second is the result of applying the conventional method. The third picture represents results for the proposed approach. As shown in the figures, we can observe that noise removal for the speech and non-speech region is remarkably performed through signals. Therefore, the proposed method shows superior noise reduction performance compared to the conventional method. However, from Figure 11, it can be seen that the speech signal is relatively small compared to other noise power levels. For this reason, as shown in Figure 6, the value of spectral energy distance between the speech region and the non-speech region is small, and the threshold value that determines the two regions is ambiguous. Therefore, it said that the noise signal of the non-speech region is reduced as much as possible, and the speech signal is also lost.

Figures 15 represent spectrogram results for the conventional method and proposed adaptive spectral subtraction approach. As shown in the figures, pictures on the first line are the original noisy speech spectrograms for each noise power level. Second and third lines represent spectrograms of the signal applied to the conventional method and the proposed approach, respectively. When comparing each of method through the figures, we can observe that the spectrogram of the proposed method has a darker background color than the spectrogram of the conventional method. This indicates the extent of noise removal, and the proposed method shows better noise reduction performance compared to the conventional method.

## 5. ORIGINALITY OF THIS WORK

The proposed method has originality and advantage compared to the conventional method. The conventional 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.

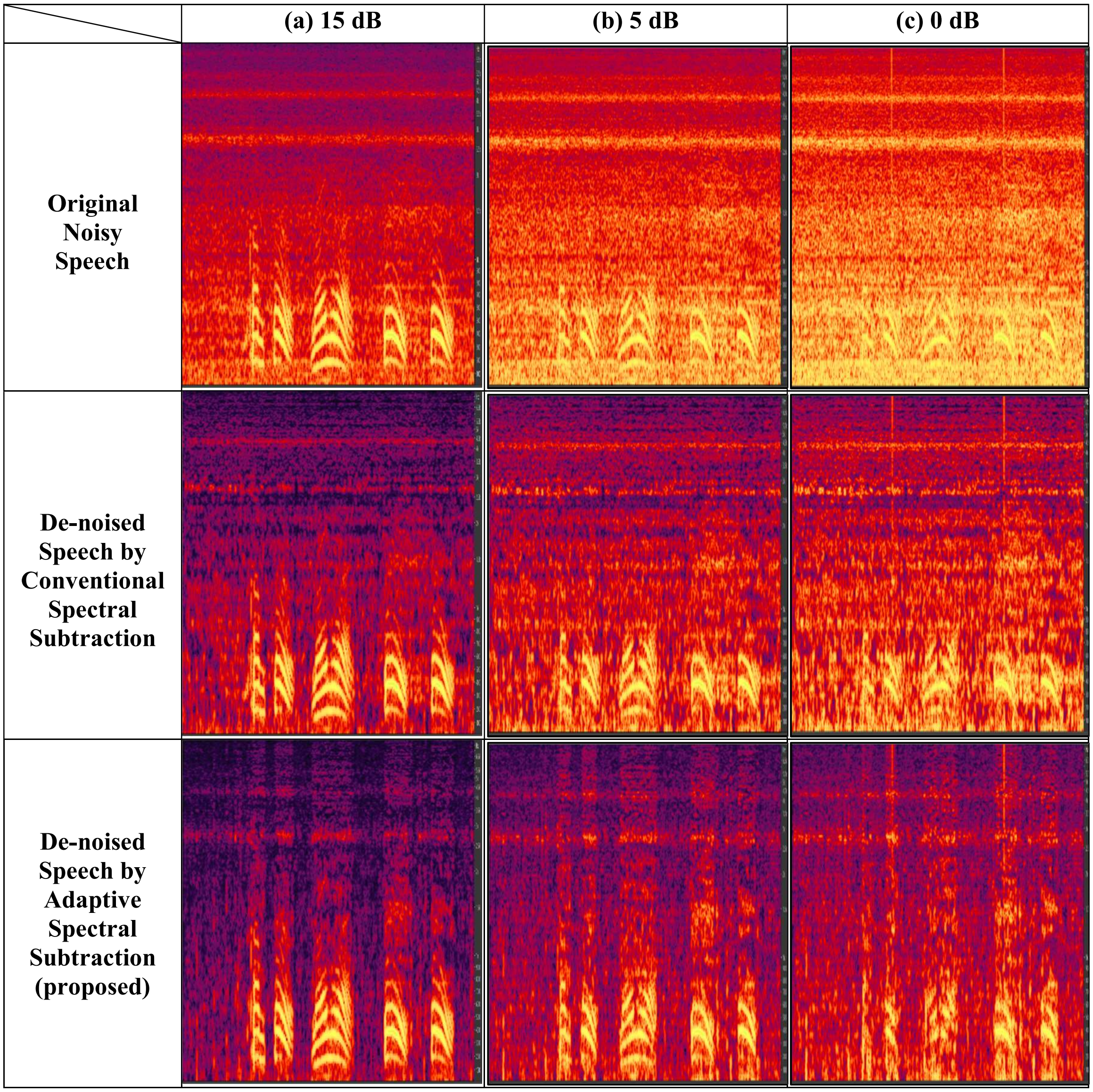

*Figure 15: Spectrograms for words, "Come to manual three two," spoken by a male speaker: The above pictures represent results for the conventional and proposed method: (a) Noisy Speech at 15 dB; (b) Noisy Speech at 5 dB; (c) Noisy Speech at 0 dB.*

The estimated spectral energy is subtracted for entire signal regions. This procedure may lead to incorrect noise reduction owing to incorrect noise estimation. In particular, the conventional method may be vulnerable to non-stationary noise signals in which noise characteristics frequently change.

On the other hand, the proposed method continuously detects non-speech regions from input contaminated signals. The spectral energy estimated from each non-speech region is used to reduce noise components in the subsequent speech region. This procedure can efficiently handle the non-stationary noise signals by frequently updating noise components.

Several experiments showed the efficiency of the proposed method compared to the conventional one. In overall noise levels, the proposed method significantly removed noise components while maintaining speech signals.

## 6. CONCLUSION

In this paper, we proposed an adaptive spectral subtraction method. The conventional spectral subtraction technique is vulnerable to non-stationary noise although it is quite effective for stationary noise. However, there are problems in that residual noise remains in non-speech regions, and noise reduction in the speech regions may not be correct. The proposed approach is to continuously update noise components of several consecutive frames and remove the components from speech regions.

We verified the noise reduction performance by conducting several experiments using noisy speech signals. The proposed method achieved a better noise reduction performance compared to the conventional approach, and enhanced speech signals were obtained.

## 7. FUTURE WORKS

In general speech recognition environment, non-stationary noise signals contaminate speech signals more frequently than stationary noise. Many researchers have hereby proposed effective noise reduction approaches such as the spectral subtraction, Wiener filer, etc. The spectral subtraction has been widely used because it has low computational complexity and is easy to implement. Our proposed method demonstrated superior noise reduction performance compared to the conventional approach. Nevertheless, it is necessary to compare the performance to other noise reduction methods for further verification. In addition, instead of using predetermined threshold, to define the threshold used for VAD procedure automatically is required for handling severe noise signals such as musical noise. Finally, we will perform speech recognition experiments to validate if the proposed method operates well for speech recognition tasks.

## 8. ACKNOWLEDGMENTS

The research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(NRF-2017R1D1A1A09000903) and the Research Grant from the Defense Acquisition Program Administration and the Agency for Defense Development under contract(UD160054BD).

## REFERENCES:

[1] J. Li, L.Deng, Y. Gong, "An overview of noise-robust automatic speech recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* Vol. 22,No. 4, 2014, pp. 745-777.

[2] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement", *IEEE Transactions on Audio, Speech and Language Processing,* Vol. 14,No. 2, 2006, pp. 2098-2108.

[3] U. Shrawankar, VM. Thakare, "Adverse conditions and asr techniques for robust speech user interface", *IJCSI International Journal of Computer Science Issues,* Vol. 8,No. 3, 2011, pp. 440-449.

[4] Y. Lu, PC. Loizou, "A geometric approach to spectral subtraction", *Speech communication,* Vol. 50,No. 6, 2008, pp. 453-466.

[5] J. Benesty, J. Chen, YA. Huang, and S. Doclo, "Study of the Wiener filter for noise reduction", *Speech Enhancement,* Springer Berlin Heidelberg, 2005, pp. 9-41.

[6] 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, "Speech enhancement with an adaptive Wiener filter", *International Journal of Speech Technology,* Vol. 17,No. 1, 2014, pp. 53-64.

[7] Y. Ephraim, D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", *EEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. 32,No. 6, 1984, pp. 1109-1121.

[8] B. Schwerin, K. Pailwal, "Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement", Speech Communication, Vol. 58, 2014, pp. 49-68

[9] B. Widrow, S.D. Stearns, "Adaptive Signal Processing", *Prentice Hall*, 1985

[10] S. So, KK. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement", *Speech Communication,* Vol. 53,No. 6, pp. 818-829

[11] C.S Ahn, .H. Choi, "Noise Reduction Algorithm using Average Estimator Least Mean Square Filter of Frame Basis", *Journal of Digital Convergence,* Vol. 11,No. 7, 2013, pp. 135-140.

[12] N Upadhyay, A Karmakar, "A Perceptually Motivated Multi-Band Spectral Subtraction Algorithm for Enhancement of Degraded Speech", *Computer and Communication Technology (ICCCT) 2012 Third international Conference on IEEE,* 2012, pp. 340-345.

[13] S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. on Acoustics, Speech and Signal Processing,* Vol. 27, No. 2, 1979, pp. 113-120.

[14] S. Yiming, W. Rui, “Voice Activity Detection Based on the Improved Dual-Threshold Method”, *IEEE International Conference on Intelligent Transportation in Big Data and Smart City (ICITBS) (2015),* 2015, pp. 996-999.