Regular paper | Journal of the KIECS. pp. 987-994, vol. 15, no. 6, Dec. 31. 2020, t. 104, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2020.15.6.987

CNN 잡음 감쇠기에서 커널 사이즈의 최적화

이행우*

Optimization of the Kernel Size in CNN Noise Attenuator

Haeng-Woo Lee*

요 약

본 논문은 음향잡음감쇠기에서 CNN(: Convolutional Neural Network) 계층의 커널 사이즈가 성능에 미치는 영향을 위한 연구하였다 이 시스템은 기존의 적응필터를 이용하는 대신 신경망 적응예측필터를 이용한 심층학습 알고리즘으로 잡음감쇠 성능을 개선한다. 100-neuron, 16-filter CNN 필터와 오차 역전파(back propagation) 알고리즘을 이용하여 잡음이 포함된 단일입력 음성신호로부터 음성을 추정한다. 이는 음성신호가 갖는 유성음 구간에서의 준주기적 성질을 이용하는 것이다. 본 연구에서 커널 사이즈에 대한 잡음감쇠기의 성능을 검증하기 위하여 Tensorflow와 Keras 라이브러리를 사용한 시뮬레이션 프로그램을 작성하고 모의실험을 수행하였다. 모의실험 결과, 커널 사이즈가 16 정도일 때 평균자승오차(MSE: Mean Square Error) 및 평균절대값오차(MAE: Mean Absolute Error) 값이 가장 작은 것으로 나타났으며 사이즈가 이보다 더 작거나 커지면 MSE 및 MAE 값이 증가하는 것을 볼 수 있다. 이는 음성신호의 경우 커널 사이즈가 16 정도일 때 특성을 가장 잘 포집할 수 있음을 알 수 있다.

ABSTRACT

In this paper, we studied the effect of kernel size of CNN layer on performance in acoustic noise attenuators. This system uses a deep learning algorithm using a neural network adaptive prediction filter instead of using the existing adaptive filter. Speech is estimated from a single input speech signal containing noise using a 100-neuron, 16-filter CNN filter and an error back propagation algorithm. This is to use the quasi-periodic property in the voiced sound section of the voice signal. In this study, a simulation program using Tensorflow and Keras libraries was written and a simulation was performed to verify the performance of the noise attenuator for the kernel size. As a result of the simulation, when the kernel size is about 16, the MSE and MAE values are the smallest, and when the size is smaller or larger than 16, the MSE and MAE values increase. It can be seen that in the case of an speech signal, the features can be best captured when the kernel size is about 16.

키워드

Noise Reduction, Deep Learning, Convolutional Neural Network, Kernel Size
잡음 감쇠, 심층 학습, CNN, 커널 크기

* 교신저자: 남서울대학교 정보통신공학과
• 접 수 일 : 2020. 09. 13
• 수정완료일 : 2020. 10. 30
• 게재확정일 : 2020. 12. 15
• Received : Sep. 13, 2020, Revised : Oct. 30, 2020, Accepted : Dec. 15, 2020
• Corresponding Author : Haeng-Woo Lee
Dept. of Information Communication Engineering, Namseoul University,
Email : hwlee@nsu.ac.kr

Ⅰ. 서 론

음성신호에 포함된 잡음을 감쇠시키는 음성개선기술에 대해 지금까지 많은 연구가 이루어지고 있다. 잡음 경감을 위한 기술은 크게 두가지 종류로 분류할 수 있다. 첫째, 짧은 구간의 스펙트럼 추정에 기반을 둔 스펙트럼 감산법[1,2]과 Wiener 필터방법[3,4,5]이 있다. 이 방법들은 추정된 잡음의 스펙트럼을 입력 음성신호에서 감산하거나 깨끗한 음성 스펙트럼을 추정하며, 잡음과 구하는 음성신호의 통계적 특성을 알고 있을 때 적합하다. 둘째는 음성신호의 준주기적 특성을 이용하는 Comb 필터[6]와 적응 필터방법[7,8,9]이 있다. Comb 필터방법은 잡음이 특정 주파수대역을 가지고 있을 때 사용되며, 적응 필터방법은 필터의 계수를 자동적으로 조정하는 기능을 가지고 있어 잡음의 통계적 특성을 미리 알고 있지 않아도 된다.

적응 잡음감쇠기는 음향센서의 수에 따라 단일입력과 다중입력 시스템으로 구분되는데 단일입력시스템[10]은 하나의 마이크를 통해서 음성신호가 입력된다. 음성신호의 유성음 구간이 나타내는 준주기적 특성을 이용하면 잡음이 포함된 마이크 입력신호로부터 음성신호를 추정할 수 있다.

딥러닝은 신경망을 기반으로 많은 수의 은닉층을 사용하는 복잡한 머신러닝 모델이다. 다층 신경망을 학습시키는 오차 역전파(back propagation) 알고리즘을 사용함으로써 많은 층으로 구성된 심층 신경망도 학습이 가능하게 되었다[11,12]. 현재 가장 많이 사용되는 딥러닝 모델은 CNN[13]이다. 본 연구에서는 적응 잡음감쇠기의 적응필터 대신에 CNN 신경망 필터의 심층학습(deep learning) 알고리즘을 이용하여 잡음을 감쇠시킬 때 커널(kernel) 사이즈가 성능에 미치는 영향을 조사하였다.

논문의 내용은 Ⅱ절에서 적응 잡음 감쇠기에 대해 알아보고, Ⅲ절에서는 CNN 신경망 필터의 구조를 설명하였으며, Ⅳ절에서는 딥러닝 모델의 역전파 알고리즘을 기술하였다. 그리고 Ⅴ절에서 커널 사이즈에 대한 시뮬레이션 및 그 결과에 대하여 기술하였고, 끝으로 Ⅵ절에서 결론을 도출하였다.

Ⅱ. 적응 잡음 감쇠기

그림 1은 음성신호의 준주기적 특성을 이용하여 적응 예측방법으로 1 샘플 이상 지연된 신호들로부터 현재 음성을 추정하는 단일입력 적응 잡음감쇠기이다. 한 두 피치 지연된 입력신호는 음성신호 성분과 높은 상관관계를 갖지만 잡음 성분과는 거의 상관관계가 없다. 따라서 음성신호는 잡음과 서로 독립된 관계이며 목표값의 최소 자승오차가 되도록 수렴해나간다.

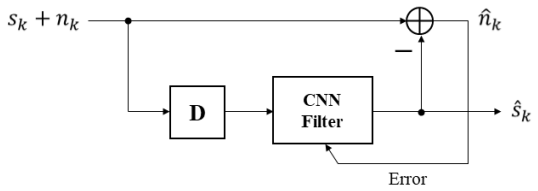

그림 1. 적응 잡음 감쇠기
Fig. 1 Adaptive noise attenuator

CNN 필터의 출력은 입력신호에 포함된 음성신호의 유성음을 추정하며 이 신호가 입력신호로부터 차감되면 잡음신호 추정값이 된다. 잡음추정신호는 CNN 필터의 계수 조정을 위한 오차신호로 사용되며 이 오차신호의 평균전력은 식 (1)과 같다.

$$E\{\hat{n_k^2}\} = E\{(s_k - \hat{s}_k)^2 + 2(s_k - \hat{s}_k)n_k + n_k^2\} \quad (1)$$

여기서 $E\{\cdot\}$는 평균값을 나타내며 음성신호와 잡음은 서로 상관관계가 없다고 가정하면

$$E\{\hat{n_k^2}\} = E\{(s_k - \hat{s}_k)^2 + n_k^2\} \quad (2)$$

임의의 프레임에서 잡음에너지는 고정된 값이므로

$$\min\left(E\{\hat{n_k^2}\}\right) = \min\left(E\{(s_k - \hat{s}_k)^2\}\right) + E\{n_k^2\} \quad (3)$$

즉 $E\{\hat{n_k^2}\}$를 최소화시키는 것은 음성신호의 추정오차 $E\{(s_k - \hat{s}_k)^2\}$를 최소화시키는 것이다. 이때 필터의 출력인 음성신호의 추정치 $\hat{s}_k$는 음성신호를 가장 잘 추정하게 된다. 따라서 $E\{(s_k - \hat{s}_k)^2\}$의 최소

화는 $E\{(n_k-\hat{n}_k)^2\}$를 최소화하는 것을 의미하며 오차신호 $\hat{n}_k$는 잡음을 추정하게 된다.

음성신호와 잡음이 혼합된 마이크 입력신호는 유성음 구간에서 준주기적 특성을 갖게 되므로 한 두 피치 지연된 신호는 음성신호와 높은 상관도를 가진다. 이때 필터의 출력은 오차신호의 에너지를 최소화함으로써 입력신호 내의 음성신호와 최소 자승오차를 갖는 음성추정신호가 된다.

III. CNN 신경망 필터의 구조

본 논문에서 사용한 그림 2의 신경망 필터는 CNN 층을 이용한 3층 구조로 되어 있다.

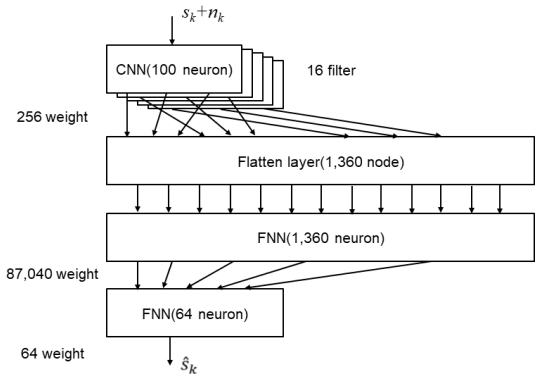

그림 2. CNN 필터의 구조
Fig. 2 Structure of CNN filter

첫째 층의 CNN 레이어는 100개 뉴런과 16개 특징 필터로 이루어져 있으며, 커널(kernel)의 크기는 16 샘플로서 매 샘플 간격으로 커널이 존재한다. 입력신호는 매 샘플마다 100×16개의 데이터로 구성되며, 출력에서 활성화(activation) 함수로 ReLU를 적용한다. CNN 층의 출력은 다음에 있는 Flatten 레이어를 거쳐 1차원으로 평탄화되며 85×16 = 1,360개 노드로 펼쳐진다. 이 신호들이 1,360개 뉴런을 갖고 있는 Fully-connected Neural Network(: FNN) 층으로 입력되며 출력에서 다시 ReLU 함수를 적용한다. 이어서 마지막 층인 64개 뉴런을 가진 FNN 레이어를 거쳐 1개의 신호로 출력된다. 계산량을 줄이기 위해 배치(batch) 크기는 32로 설정하였으며 각 레이어의 bias 파라미터는 생략하였다. 이 모델에서 계산해야 할 가중치 파라미터는 CNN 레이어에서 256개(=16×16), 은닉 레이어에서 87,040개(=1,360×64), 그리고 출력 레이어에서 64개로서 총 87,360개이다. 가중치의 업데이트 알고리즘으로 Adam과 오차 역전파 알고리즘을 이용한다. 본 시스템은 지도학습으로 분류되므로 단일입력 데이터로 훈련데이터와 학습 목표값을 준비한다.

IV. 딥러닝 모델의 역전파 알고리즘

다층 퍼셉트론(multi-layer perceptron)은 한 개 이상의 은닉층(hidden layer)을 가지는 다층 신경망의 구조를 갖고 있다. 그림 3은 l개의 입력 뉴런(neuron)을 가지는 입력층, m개의 은닉 뉴런을 가지는 은닉층, n개의 출력 뉴런을 가지는 출력층으로 구성된 다층 퍼셉트론을 나타낸다. 다층 퍼셉트론의 입력 뉴런의 값은 l차원 벡터 $x=[x_1,x_2,\cdots,x_i,\cdots,x_l]$로 나타내고, 은닉 뉴런의 값은 m차원 벡터 $a^1=[a_1^1,a_2^1,\cdots,a_j^1,\cdots,a_m^1]$로 나타내며, 출력 뉴런의 값은 n차원 벡터 $y=[y_1,y_2,\cdots,y_k,\cdots,y_n]$로 표현한다. 그리고 입력층과 은닉층 사이의 가중치(weight)를 w_{ij}^1, 은닉층과 출력층 사이의 가중치를 w_{jk}^2로 나타내고 바이어스(bias)는 생략해도 된다. 여기서 입력층은 CNN 레이어로서 커널(kernel) 사이즈 q와 필터 수 p로 이루어져 있다. 또한, j번째 은닉 뉴런으로 입력되는 가중합을 u_j^h, k번째 출력 뉴런으로 입력되는 가중합을 u_k^o라 하고, 은닉 뉴런의 활성화 함수는 식 (4)의 ReLU 함수를 사용하며 ϕ_{relu}로 표기하고 출력 뉴런은 활성화 함수를 사용하지 않는다.

$$\begin{aligned}\phi_{relu}(z) &= z \quad for \quad z>0 \\ &= 0 \quad for \quad z\le 0\end{aligned} \tag{4}$$

그러면 은닉 뉴런과 출력 뉴런의 출력값은 식 (5)와 (6)으로 나타낼 수 있다.

$$a_j^1 = \phi(u_j^h) = \phi_{relu}\left(\sum_{i=1}^{l} w_{ij}^1 x_i\right) \quad (5)$$

$$y_k = u_k^o = \sum_{j=1}^{m} w_{jk}^2 a_j^1 \quad (6)$$

모든 가중치를 하나의 파라미터 θ로 나타내면 입력 x가 주어졌을 때 k번째 출력 뉴런의 값은 함수 $f_k(x,\theta)$로 표현된다.

$$f_k(x,\theta) = y_k = \sum_{j=1}^{m} w_{jk}^2 \phi_{relu}\left(\sum_{i=1}^{l} w_{ij}^1 x_i^1\right) \quad (7)$$

오차 역전파 학습알고리즘[14]은 다층 퍼셉트론을 학습시킬 수 있는 알고리즘으로 다층 퍼셉트론의 지도학습은 학습목표 출력값이 주어지고 다층 퍼셉트론에 의해 출력되는 값의 차이를 이용한 오차함수가 정의되어야 한다. 학습데이터와 목표 출력값이 입출력의 순서쌍 (x_i, t_i)로 주어졌을 때 학습 데이터 전체 X에 대한 오차는 식 (8)과 같이 평균제곱오차로 정의할 수 있다.

$$E(X,\theta) = \frac{1}{2N}\sum_{i=1}^{N} \| t_i - f(x_i,\theta) \|^2 \quad (8)$$

위 식에서 오차함수 $E(X,\theta)$는 데이터 집합 X와 파라미터 θ가 주어지면 하나의 값으로 정해지는데 X는 외부에서 주어지는 값이고 최적화해야 하는 대상은 θ이므로 일반적으로 $E(\theta)$로 나타낼 수 있다. 오차 역전파 학습 알고리즘은 오차함수 $E(\theta)$를 최소화하기 위한 파라미터를 찾기 위해 경사하강법(gradient descent method)을 사용한다. 경사하강법은 어떤 비용함수의 값을 최소화시키는 파라미터를 반복적 탐색으로 찾는 알고리즘으로 식 (9)로 표현된다.

$$\theta(t+1) = \theta(t) + \Delta\theta(t) = \theta(t) - \eta\frac{\partial E(\theta)}{\partial \theta} \quad (9)$$

여기서 η는 학습의 속도를 조절하는 학습률(learning rate)이다. 다층 퍼셉트론에서 오차 역전파 학습은 각 가중치마다 하나의 데이터를 사용하여 업데이트하는 확률(stochastic) 경사하강법을 적용하며 하나의 데이터에 대한 오차함수 $E(x,\theta)$를 사용한다.

$$E(x,\theta) = \frac{1}{2}(t_k - y_k)^2 = \frac{1}{2}\left(t_k - \sum_{j=1}^{m} w_{jk}^2 a_j^1\right)^2 \quad (10)$$

여기서 학습을 통해 수정해야 하는 파라미터는 은닉층과 출력층 사이의 가중치 w_{jk}^2와 입력층과 은닉층 사이의 가중치 w_{ij}^1이다. 먼저 오차함수를 출력층 가중치로 편미분하면

$$\frac{\partial E}{\partial w_{jk}^2} = \frac{\partial E}{\partial u_k^o}\frac{\partial u_k^o}{\partial w_{jk}^2} = -(t_k - y_k)a_j^1 = \delta_k a_j^1 \quad (11)$$

여기서 δ_k는 출력 뉴런이 오차에 미치는 영향이다. 오차함수를 입력층 가중치로 편미분하면

$$\frac{\partial E}{\partial w_{ij}^1} = \frac{\partial E}{\partial u_j^h}\frac{\partial u_j^h}{\partial w_{ij}^1} = \phi_h^{'}(u_j^h)\sum_{j=1}^{m} w_{jk}^2 \delta_k x_i = \delta_j x_i \quad (12)$$

종합해보면 입력층과 은닉층 사이의 가중치는 은닉층과 출력층 사이의 가중치와 각각의 출력 뉴런이 오차에 미치는 영향인 δ_k를 곱하여 합한 값에 영향을 받는 것을 알 수 있다. 이처럼 출력 뉴런의 오차가 은닉 뉴런에 거꾸로 전파되어 은닉 뉴런의 파라미터 조절에 영향을 미치기 때문에 다층 퍼셉트론의 경사하강 학습법을 오차 역전파 학습알고리즘이라 하며 최종적으로 각 가중치의 업데이트는

$$w_{jk}^2(t+1) = w_{jk}^2(t) + \eta(t_k - y_k)a_j^1 \quad (13)$$

$$w_{ij}^{1}(t+1)=w_{ij}^{1}(t)-\eta\phi_{h}^{'}(u_{j}^{h})\sum_{j=1}^{m}w_{jk}^{2}\delta_{k}x_{i} \quad (14)$$

Ⅴ. 모의실험 결과

본 논문에서 제안한 음성잡음감쇠기의 성능을 검증하기 위해 Tensorflow와 Keras 라이브러리를 이용하여 시뮬레이션 프로그램을 작성하였다. 입력신호는 음성과 백색잡음이 혼합되어 8kHz로 샘플링되며, 500,000 샘플(62.5 sec)을 준비하였다. 이 시스템은 지도학습에 해당되므로 입력데이터는 내부적으로 100×499,901 샘플의 입력배열과 499,901 샘플의 목표값으로 구성된다.

그림 3은 잡음이 혼합된 입력 음성신호의 파형을 나타낸다.

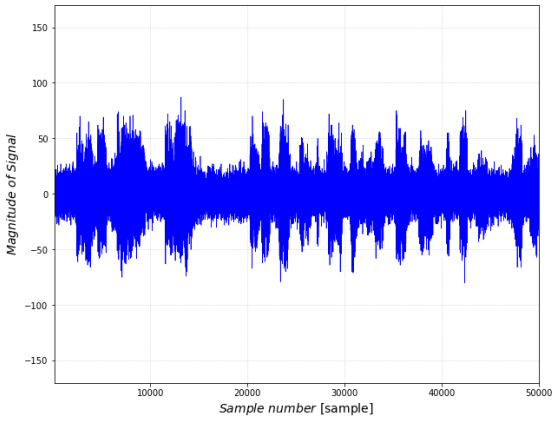

그림 3. 잡음이 혼합된 입력 음성신호
Fig. 3 Input speech signal with noises

시스템 간의 성능을 평가하기 위하여 목표값인 입력신호와 음성예측값 간의 오차에 대한 평균자승오차 MSE와 평균절대값오차 MSE를 사용하였다.

그림 4에서 4개 커널 사이즈(8, 12, 16, 20) 별 MSE 곡선을, 그림 5에서도 4개 커널 사이즈(8, 12, 16, 20) 별 MAE 곡선을 비교하였다. 이 두 그림으로부터 커널 사이즈가 커질수록 MSE 및 MAE가 모두 비슷하게 감소하는 것을 볼 수 있다.

또한 그림 6에서는 6개 커널 사이즈(4, 8, 12, 16, 20, 24)에 대한 MSE 곡선을, 그리고 그림 7에서는 6개 커널 사이즈(4, 8, 12, 16, 20, 24)에 대한 MAE 곡선을 도시하였다.

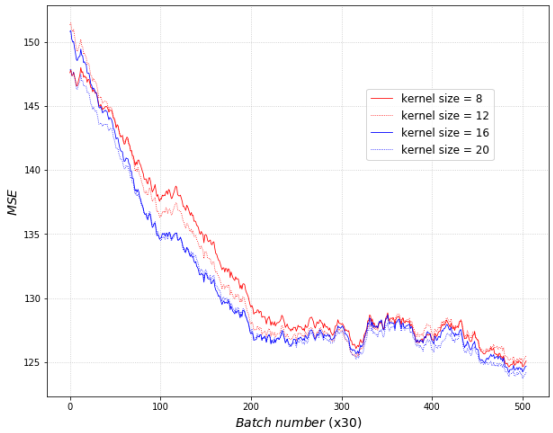

그림 4. 커널 크기별 평균자승오차의 비교
Fig. 4 Comparison of MSE for kernel sizes

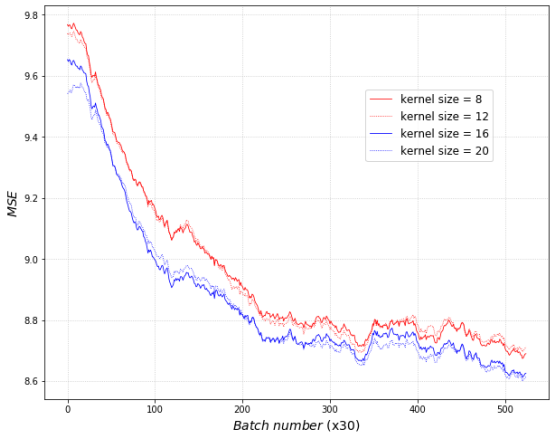

그림 5. 커널 크기별 평균절대값오차의 비교
Fig. 5 Comparison of MAE for kernel sizes

그림 6에서는 커널 사이즈가 16일 때 MSE가 가장 작게 나타나고 이를 중심으로 사이즈가 작아지거나 커지면 MSE가 증가하는 것으로 나타났다. 이와 함께 그림 7에서도 커널 사이즈가 16일 때 MAE가 가장 작고 사이즈가 작거나 큰 경우에는 MAE가 커지는 것을 볼 수 있다. 이는 음성신호의 특성이 커널 사이즈가 16 정도일 때 가장 잘 표출된다는 것을 알 수 있다.

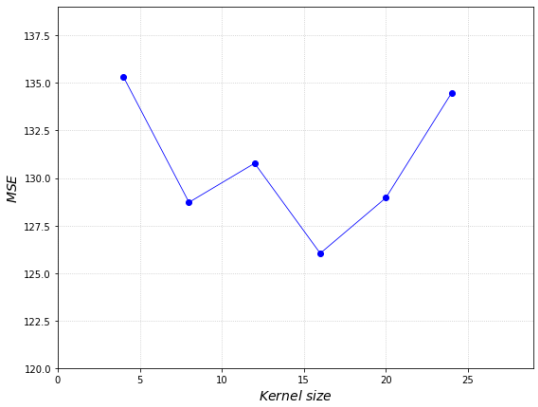

그림 6. 커널 크기에 대한 평균자승오차 곡선
Fig. 6 MSE curve to kernel sizes

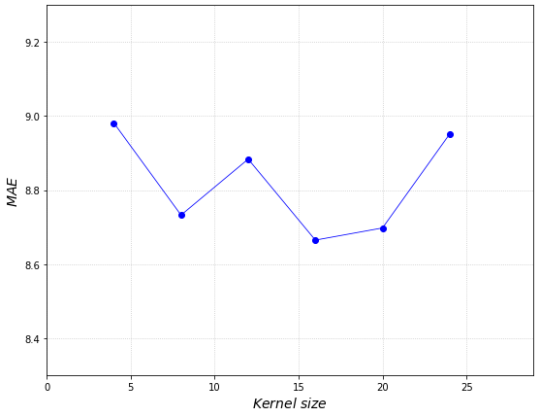

그림 7. 커널 크기에 대한 평균절대값오차 곡선
Fig. 7 MAE curve to kernel sizes

Ⅵ. 결 론

본 논문에서는 CNN 딥러닝 기술을 이용한 잡음감쇠기에서 커널 사이즈가 성능에 미치는 영향을 살펴보았다. 잡음감쇠기는 100-neuron, 16-filter CNN 필터와 오차 역전파 알고리즘을 이용하여 구현하였다.

Tensorflow 및 Keras 라이브러리를 사용하여 모델을 코딩하였고 커널 사이즈에 따라 MSE 및 MAE 값이 어떻게 변화하는지 관찰하였다. 모의실험 결과, 본 시스템은 커널 사이즈가 16일 때 MSE 및 MAE 값이 가장 작은 것으로 나타났고 이를 중심으로 사이즈가 작아지거나 커지면 MSE 및 MAE 값이 증가하는 것을 보여주었다. 이것은 음성신호의 경우 커널 사이즈가 16 정도일 때 특성을 가장 잘 포착할 수 있기 때문이다.

References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, Apr. 1979, pp. 113-120.

[2] A. Schaub and P. Schaub, "Spectral sharpening for speech enhancement/noise reduction," *Proc. of Int. Conf. on Acoust., Speech, Signal Processing*, vol. 2, May 1991, pp. 993-996.

[3] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, Jun. 1978, pp. 197-210.

[4] J. Hansen and M. Clements, "Constrained iterative speech enhancement with to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-39, no. 4, Apr. 1989, pp. 21-27.

[5] J. Choi, "Noise Reduction Algorithm in Speech by Wiener Filter," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 8, Sep. 2013, pp. 1293-1298.

[6] J. S. Lim, A. V. Oppenheim and L. D. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, no. 4, Apr. 1991, pp. 354-358.

[7] S. F. Boll and D. C. Pulsipher, "Suppression of acoustic noise in speech using two microphone adaptive noise cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 6, Dec. 1989, pp. 752-753.

[8] W. A. Harrison, J. S. Lim, and E. Singer, "A new application of adaptive noise cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.

ASSP-34, Feb. 1986, pp. 21-27.

[9] O. S. Kwon, "Study on Efficient Adaptive Controller for Attenuation of Engine Noises in a Car," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 9, Sep. 2014, pp. 983-989.

[10] M. R. Sambur, "Adaptive noise canceling for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, Oct. 1978, pp. 419-423.

[11] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, 2015, pp. 85-117.

[12] J. Choi, "Speech and Noise Recognition System by Neural Network," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 5, Aug. 2010, pp. 357-362.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, Nov. 1998, pp. 2278-2324.

[14] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, 1988, pp. 3.

저자 소개

이행우(Haeng-Woo Lee)

1985년 광운대학교 전자공학과 (공학사)
1987년 서강대학교 대학원 전자공학과 (공학석사)
2001년 전북대학교 대학원 전자공학과 (공학박사)
1987년～1998년 한국전자통신연구원 선임연구원
2001년～현재 남서울대학교 정보통신공학과 교수
※관심분야 : VLSI 설계, 딥러닝, 음향잡음감쇠