Regular paper | Journal of the KIECS. pp. 113-118, vol. 19, no. 1, Feb. 29. 2024, t. 123, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2024.19.1.113

DWT 기반 딥러닝 잡음소거기에서 웨이블릿 최적화

정원석* · 이행우**

Optimizing Wavelet in Noise Canceler by Deep Learning Based on DWT

Won-Seog Jeong* · Haeng-Woo Lee**

요 약

본 논문에서는 음향신호의 배경잡음을 소거하기 위한 시스템에서 최적의 wavelet을 제안한다. 이 시스템은 기존의 단구간 푸리에변환(STFT: Short Time Fourier Transform) 대신 이산 웨이블릿변환(DWT: Discrete Wavelet Transform)을 수행한 후 심층학습과정을 통하여 잡음소거 성능을 개선하였다. DWT는 다해상도 대역통과필터 기능을 하며 각 레벨에서 모 웨이블릿을 시간 이동시키고 크기를 스케일링한 여러 웨이블릿을 이용하여 변환 파라미터를 구한다. 여기서 음성을 분석하는데 가장 적합한 모(mother) 웨이블릿을 선정하기 위해 여러 웨이블릿에 대한 잡음소거 성능을 실험하였다. 본 연구에서 여러 웨이블릿에 대한 잡음소거시스템의 성능을 검증하기 위하여 Tensorflow와 Keras 라이브러리를 사용한 시뮬레이션 프로그램을 작성하고 가장 많이 사용되는 4개의 wavelet에 대해 모의실험을 수행하였다. 실험 결과, Haar 또는 Daubechies 웨이블릿을 사용하는 경우가 가장 우수한 잡음소거 성능을 나타냈으며 타 웨이블릿을 사용하는 경우보다 평균자승오차(MSE: Mean Square Error)가 크게 개선되는 것을 볼 수 있었다.

ABSTRACT

In this paper, we propose an optimal wavelet in a system for canceling background noise of acoustic signals. This system performed Discrete Wavelet Transform(DWT) instead of the existing Short Time Fourier Transform(STFT) and then improved noise cancellation performance through a deep learning process. DWT functions as a multi-resolution band-pass filter and obtains transformation parameters by time-shifting the parent wavelet at each level and using several wavelets whose sizes are scaled. Here, the noise cancellation performance of several wavelets was tested to select the most suitable mother wavelet for analyzing the speech. In this study, to verify the performance of the noise cancellation system for various wavelets, a simulation program using Tensorflow and Keras libraries was created and simulation experiments were performed for the four most commonly used wavelets. As a result of the experiment, the case of using Haar or Daubechies wavelets showed the best noise cancellation performance, and the mean square error(MSE) was significantly improved compared to the case of using other wavelets.

키워드
Deep learning, Discrete wavelet transform, Fully-connected neural network, Noise attenuator
심층 학습, 이산 웨이블릿 변환, FNN, 잡음 소거기

* ㈜AICube
** 교신저자 : 남서울대학교 지능정보통신공학과
• 접 수 일 : 2023. 11. 16
• 수정완료일 : 2023. 12. 30
• 게재확정일 : 2024. 02. 17
• Received : Nov. 16, 2023, Revised : Dec. 30, 2023, Accepted : Feb. 17, 2024
• Corresponding Author : Haeng-Woo Lee
Dept. Intelligent Information and Communication Engineering, Namseoul University
Email : hwlee@nsu.ac.kr

Ⅰ. 서 론

배경잡음 소거기술은 음성신호에서 배경잡음을 소거하여 화자의 명료한 음성을 청취할 수 있도록 하는 것으로, 보청기나 스마트폰뿐만 아니라 오래된 음반의 음질 향상 혹은 음성인식 시스템까지 다양한 분야에서 활용되고 있다. 배경잡음을 소거하기 위한 대표적인 기술로는 입력되는 음성신호의 짧은 구간에 대한 스펙트럼 추정에 기반을 둔 스펙트럼 감산법[1]과 위너(wiener) 필터방법[2][3]이 있다. 이러한 스펙트럼 감산법과 위너 필터방법은 추정된 배경잡음의 스펙트럼을 입력 음성신호에서 감산하거나, 깨끗한 음성 스펙트럼을 추정하는 것으로 배경잡음과 음성신호의 통계적인 특성을 알고 있을 때 적합한 기술이다. 신호의 통계적인 특성을 알지 못하는 경우에 음성신호로부터 배경잡음을 소거하는데 이용되는 기술로는 음성신호의 준주기적 특성을 이용하는 콤(comb) 필터방법[4]과 적응필터 방법[5]이 있다. 콤 필터 방법은 배경잡음이 특정 주파수대역을 가지고 있을 때 사용되고, 적응필터 방법은 필터의 계수를 자동적으로 조정하는 기능을 통해 배경잡음을 경감 또는 소거하는 것으로 배경잡음의 통계적 특성을 미리 알고 있지 않아도 된다. 그러나 음성신호에 포함되는 배경잡음은 군중 소음, 식당 대화음, 도로의 차량 통행음, 열차운행 소음, 차량 경적음, 기계 작동음 등 그 형태가 매우 다양하고 음성과 유사한 특징을 가질 수 있으며 예측 불가능하게 불규칙적으로 생성될 수 있다. 즉, 음성신호는 비정재성(non-stationary)을 가진다. 따라서 배경잡음이 특정 주파수 대역을 가지고 있을 때, 배경잡음을 소거하는 콤 필터 방법은 배경잡음이 고정된 특정 환경에서 사용할 수 있으나 실제 환경에서 적용하기에 한계가 있다. 일반적으로 시간에 따라 변하는 신호의 주파수 해석을 위해 푸리에 변환(fourier transform)이 이용되고 있으나[6] 음성신호와 같이 비정재성을 갖는 신호의 특성을 표현하기에는 적합하지 않은 면이 있다. 이러한 문제점을 극복하고자 최근에 널리 연구되고 있는 웨이블릿변환은 다중 해상도(multi resolution)를 갖는 신호해석방법[7][8]으로 시간 및 주파수의 국부성을 가지므로 신호의 통계적 특성을 알지 못하거나, 시간적으로 예측하기 어려운 신호해석에 유용하다. 이러한 웨이블릿 변환은 신호의 저주파 부분에서 좋은 주파수 분해능과 고주파 부분에서 좋은 시간 분해능을 얻을 수 있는 장점이 있다. 이로 인해 웨이블릿변환을 이용한 특징벡터를 이용하면 파열음이나 마찰음에서와 같이 시간-주파수 상에서 갑자기 튀는 국부적 특성을 잘 반영할 수 있기 때문에 음성신호를 해석하는데 적합한 방법이다. 최근에는 필터뱅크 방식을 이용한 웨이블릿 LMS 필터를 구성하여 잡음 소거기에서 기준신호가 있는 경우 빠른 속도로 배경잡음을 소거할 수 있는 기술이 제안되었으나, 문턱치(threshold) 기법[9]을 적용하기 어렵고 입력신호의 특성을 쉽게 파악할 수 없는 단점이 있다.

따라서 본 논문에서는 적응필터를 사용하지 않고 이산웨이블릿변환과 딥러닝 모델[10][11]을 이용하여 음성신호와 배경잡음신호를 포함하는 혼합신호로부터 배경잡음을 소거하는 방안을 제안하고자 한다. 즉, 본 논문은 배경잡음과 음성신호를 포함하는 혼합신호의 이산웨이블릿변환 계수를 학습하여 혼합신호에 포함된 순수한 음성신호에 대한 이산웨이블릿변환 계수를 추정하기 위해 생성한 딥러닝 모델에 실제 혼합신호를 이산웨이블릿변환하여 획득한 이산웨이블릿변환 계수를 적용함으로써 실제 혼합신호에 포함된 순수 음성신호의 이산웨이블릿변환 계수를 추정하고, 추정한 순수 음성신호의 이산웨이블릿변환 계수에 따라 역이산웨이블릿변환을 수행하여 순수한 음성신호를 추정함으로써 배경잡음을 효과적으로 소거하는 방안을 제안하고자 한다.

논문의 내용은 Ⅱ절에서 이산 웨이블릿 변환에 대해 살펴보고, Ⅲ절에서는 잡음제거를 위한 변환영역 블록 딥러닝 모델을 제안하였다. 그리고 Ⅳ절에서 이 시스템에 대한 시뮬레이션 및 그 결과에 관하여 기술하였고, 끝으로 Ⅴ절에서 결론을 도출하였다.

Ⅱ. 이산 웨이블릿 변환

웨이블릿 변환은 기저함수들의 집합에 의한 신호의 분해로서 이해할 수 있다. 이때 웨이블릿 변환에서 하나의 기저함수를 웨이블릿이라 하며 웨이블릿은 하나의 대역통과필터라고 할 수 있다. 모든 주파수에 대해 균일한 시간 분해능을 제공하는 STFT와 달리 웨이블릿 변환은 고주파수에 대해서는 높은 시간 분해능

과 낮은 주파수 분해능을 제공하고 저주파수에 대해서는 높은 주파수 분해능과 낮은 시간 분해능을 제공한다. 이는 유사한 시간-주파수 분해능 특성을 나타내는 인간의 귀와 매우 흡사하다.

웨이블릿 변환은 모 웨이블릿을 시간축으로 이동(shifting)시키고 스케일링(scaling)한 여러 웨이블릿 기저(basis)들을 이용하여 신호를 분석한다. 다해상도 신호해석에서 주어진 함수는 다른 해상도를 가진 연속적인 추정치들의 합계로써 표현되며 스케일링 함수라고 하는 저주파 통과 커널과 콘볼루션함으로써 이루어진다. DWT는 식 (1)로 정의되며 시간 이동 및 스케일링 파라미터가 이산적인 값을 갖는다. 이 식에서 ψ^*는 원형 웨이블릿이며, 이를 신호의 주파수에 따라 스케일링 파라미터 j와 시간 이동 파라미터 k에 의해 변형되어 적용된다. 즉, 푸리에 변환과 같이 고정된 크기의 창함수를 사용하지 않고 짧은 지속시간을 갖는 고주파 신호에 대해서는 짧은 창함수를 사용하고 긴 지속시간을 갖는 저주파신호에 대해서는 긴 창함수를 이용함으로써 주파수 영역에 따른 다중 해상도를 갖는다.

$$y[n] = \sum_{k=0}^{K-1} x[k] \cdot \psi^*[n-k] \quad \cdots (1)$$

실제 응용에서는 a와 b를 2의 지수 형태로 나타낸 dyadic 웨이블릿 변환(DyWT: Dyadic Wavelet Transform)이 많이 이용된다. 원 이산신호는 다해상도 분석의 다운 샘플링을 통해 주파수가 다른 여러 개의 부대역(sub-band)으로 분해되고, 업 샘플링을 통해 원 이산신호로 합성된다. 각 레벨에서 시간영역 이산신호는 저역통과필터 $H(z)$를 통해 근사(approximation) 성분과 고역통과필터 $G(z)$를 통해 상세(detail) 성분으로 분해된다.

$$\begin{aligned} x(t) &= \sum_{j,k} a_j(k)\Phi_{j,k}(t) + \sum_{j,k} d_j(k)\Psi_{j,k}(t) \quad \cdots(2) \\ &= H(z) + G(z) \end{aligned}$$

근사성분은 근사계수 $a_j(k)$와 척도함수(Scale function) $\phi(t)$의 곱으로 표현되고 상세성분은 상세계수 $d_j(k)$와 상세함수(Detail function) $\psi(t)$의 곱으로 구성되며 척도함수와 상세함수는 서로 직교한다. 여기서 j와 k는 각각 이산 변환에서의 스케일과 시간영역에서의 이동을 나타낸다.

$$\Phi_{j,k}(t) = \frac{1}{\sqrt{2^j}} \Phi\left(\frac{t-k2^j}{2^j}\right) \quad \cdots (3)$$

$$\Psi_{j,k}(t) = \frac{1}{\sqrt{2^j}} \Psi\left(\frac{t-k2^j}{2^j}\right) \quad \cdots (4)$$

각 레벨의 근사성분이 2배 간축된 출력은 그 다음 레벨의 입력신호가 된다. 각 레벨에서 근사계수는 입력신호와 척도함수가 결합되고 상세계수는 입력신호와 상세함수가 결합된 형태로 표현된다.

$$a_j(k) = \int_{-\infty}^{\infty} x(t) \cdot \phi_{j,k}(t)\, dt \quad \cdots (5)$$

$$d_j(k) = \int_{-\infty}^{\infty} x(t) \cdot \psi_{j,k}(t)\, dt \quad \cdots (6)$$

웨이블릿 변환은 신호처리 관점에서 대역통과필터뱅크의 출력으로 볼 수 있으며, 신호를 분할하기 위해 트리 형태의 웨이블릿 분해 필터뱅크를 구성한다. 입력신호가 저역통과필터와 고역통과필터를 거치고 2배의 간축(decimation) 과정을 거치면 한 번의 웨이블릿 변환이 수행되며, 이러한 과정을 원하는 스케일까지 반복적으로 수행하면 웨이블릿 변환된 신호를 얻을 수 있다. 그리고 일반적인 웨이블릿 변환에서 각 스케일은 상위 스케일에서 2배로 간축하여 구해지므로 각 스케일의 샘플 수가 상위 스케일의 절반이 된다.

III. DWT 기반 잡음 소거시스템

본 논문에서는 기존의 STFT 대신에 이산웨이블릿 변환 기술을 적용하여 잡음소거 성능을 개선하고자 한다. 변환영역 딥러닝 방법은 입력음성의 웨이블릿 변환계수를 목표값인 순수 음성의 변환계수와 같아지도록 신경망의 가중치를 조정해 나간다. 그림 1은 웨이블릿 변환을 이용하여 기준신호의 변환계수를 추정하는 적응 잡음소거시스템이다.

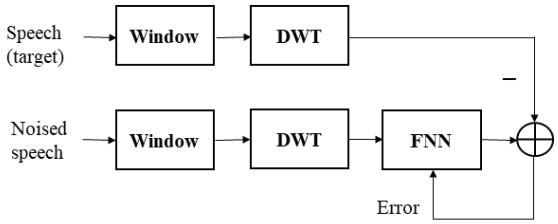


그림 1. 딥러닝기반 잡음소거시스템
Fig. 1 Noise reduction system based deep learning

이 시스템에는 잡음이 포함된 음성신호뿐만 아니라 순수한 음성신호도 입력되어 딥러닝 학습의 목표값으로 사용된다. 각 입력신호의 이산웨이블릿변환을 구하기 위해 신호를 음성의 통계적 특성이 변하지 않는 32ms 구간으로 나누고 512 샘플에 대해 해밍(Hamming) 윈도우 함수와 곱한다. 그리고 이 블록에 대해 9 레벨의 웨이블릿 변환계수를 산출하고 변환계수들을 1차원 배열로 정리한다. 이 배열이 딥러닝의 입력 데이터로 사용된다.

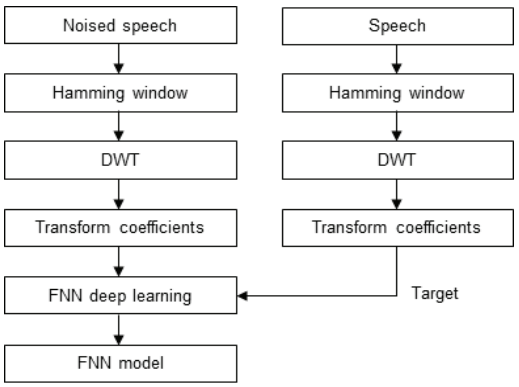


그림 2. 딥러닝 모델 생성과정
Fig. 2 Process generating deep learning model

본 논문에서 제안하는 이산웨이블릿변환을 이용한 딥러닝 기반 배경잡음 소거시스템은 화자의 음성신호와 배경잡음신호가 포함된 혼합신호를 수신하여 소정의 시간단위로 분할함으로써 복수의 구간으로 혼합신호를 분할한다. 이어서 구간별 혼합신호에 대해 창함수를 적용하고 이산웨이블릿변환을 수행하여 구간별 혼합신호의 이산웨이블릿변환 계수를 구한다. 획득한 구간별 혼합신호의 이산웨이블릿변환 계수를 딥러닝 모델에 입력하여 구간별로 혼합신호에 포함된 음성신호의 이산웨이블릿변환 계수를 추정하고 다시 구간별로 역이산웨이블릿변환을 수행함으로써 혼합신호에 포함된 음성신호를 추정하고, 딥러닝 모델 생성과정은 그림2와 같다. 혼합신호를 분할하였을 때, 각 구간의 혼합신호의 처음과 끝에서 이전 및 이후 구간의 혼합신호에 대한 연속성을 유지하기 위해서 해밍 윈도우를 적용하고 구간별 혼합신호에 대해서 9-레벨 이산웨이블릿변환을 수행하여 각 레벨별 이산웨이블릿변환 계수를 구한다. 레벨별 이산웨이블릿변환 계수는 각 레벨별로 입력신호가 고역통과필터를 통과하였을 때 산출되는 상세계수(detail coefficients)를 포함하되, 마지막 레벨에서는 저역통과필터를 통과하였을 때 산출되는 근사계수(approximation coefficients)를 포함하도록 한다. 딥러닝 모델은 복수의 학습용 혼합신호에 대한 각 구간별 이산웨이블릿변환 계수를 학습하여 생성하며, 실제 혼합신호의 각 구간별 이산웨이블릿변환 계수가 입력되면 각 구간별로 음성신호의 이산웨이블릿변환 계수를 추정하도록 구성된다. 또한 이 시스템은 각 구간별 역이산웨이블릿변환 결과를 중첩하여 결합함으로써 혼합신호에 포함된 음성신호를 출력하고, 음성 재생과정은 그림3과 같다.

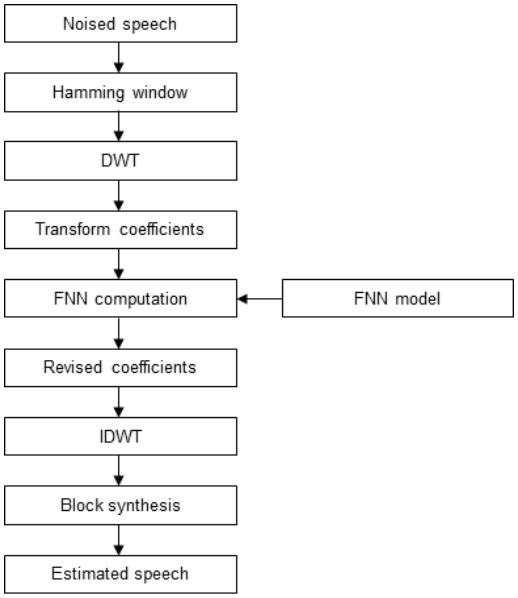
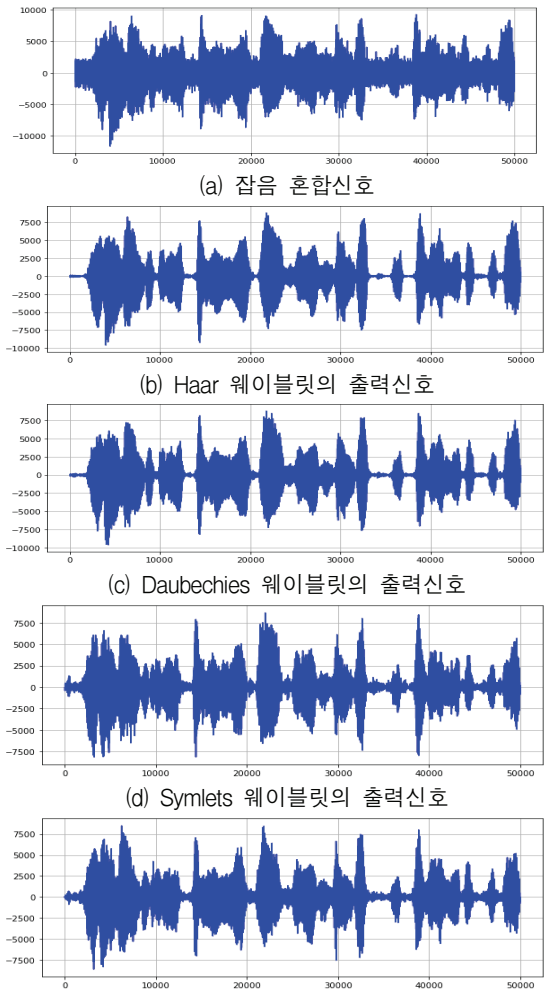


그림 3. 딥러닝 모델을 이용한 음성 재생과정
Fig. 3 Process estimating speech by using deep learning model

Ⅳ. 웨이블릿의 성능 비교분석

본 논문에서 제안한 음성잡음소거기의 성능을 검증하기 위해 Tensorflow와 Keras 라이브러리를 이용하여 시뮬레이션 프로그램을 작성하였다. 입력신호는 음성과 잡음의 혼합신호와 순수 음성신호가 사용되며 16-bit, 16kHz로 샘플링된 500,000 샘플(28.125 sec) 데이터가 제공된다. 이 시스템은 지도학습에 해당되며 입력데이터는 내부적으로 512×(500,000-511) 샘플의 입력배열과 (500,000-511) 샘플의 목표값으로 구성된다.



(a) 잡음 혼합신호

(b) Haar 웨이블릿의 출력신호

(c) Daubechies 웨이블릿의 출력신호

(d) Symlets 웨이블릿의 출력신호

(e) Coiflets 웨이블릿의 출력신호

그림 4. 각 웨이블릿에 대한 출력 파형
Fig. 4 Waveforms of output signals for wavelets

그림 4에서 4개의 wavelet에 대한 출력파형을 보여주고 있으며 가로축은 샘플 수를 나타내고 세로축은 16-bit 정수로 표현된 신호 크기를 나타낸다. 그리고 그림 5는 가로축의 배치 수에 따라서 세로축의 오차값이 감소하는 추세를 보여주는 MSE 특성곡선이다.

그림 4에서 (a)는 음성과 잡음이 혼합된 입력신호의 파형을 보여주고 (b) - (e)는 가장 많이 사용되는 4개 wavelet에 대한 출력 파형을 보여준다. 이 그림에서 볼 수 있는 바와 같이 Haar와 Daubechies 웨이블릿을 사용한 경우가 가장 잡음이 크게 소거되고 Symlets이나 Coiflets 웨이블릿을 사용하면 더 많은 잡음이 남아있는 것으로 나타났다. Haar와 Daubechies 웨이블릿은 사실상 같은 파형으로서 음향잡음을 소거하는 데에는 Haar 웨이블릿을 사용하는 것이 가장 우수한 음질개선을 기대할 수 있다.

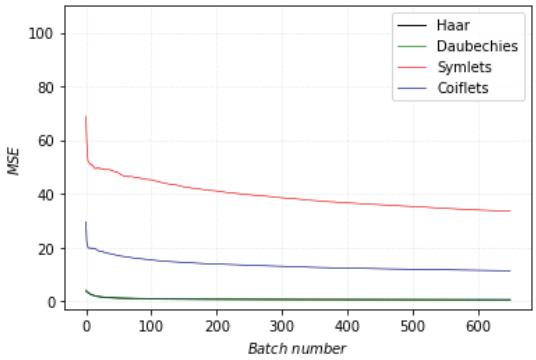


그림 5. 웨이블릿별 평균제곱오차의 특성 비교
Fig. 5 Comparison of MSE for wavelets

그림 5는 4개의 wavelet을 사용한 경우 학습시간 동안 목표값에 대한 평균제곱오차(MSE)의 변화곡선을 보여준다. 이 그림으로부터 Symlet 웨이블릿을 사용한 경우 가장 큰 오차를 나타내고 Coiflets 웨이블릿을 사용한 경우는 좀더 오차의 크기가 감소하였다. 그리고 Haar와 Daubechies 웨이블릿을 사용한 경우에는 동일하게 가장 작은 오차를 나타낸다.

Ⅴ. 결 론

배경잡음이 포함된 음향신호의 음질을 개선하기 위하여 우수한 잡음소거기의 개발이 요구되

고 있다. 본 논문에서는 웨이블릿 변환과 딥러닝 기술을 적용한 잡음소거시스템을 제안하였다. 이때 사용되는 여러 형태의 wavelet 중에서 어떤 웨이블릿을 사용하는 것이 음향신호에서 잡음을 효과적으로 소거하는지를 모의실험을 통해 검토하였다. 가장 많이 사용되는 4개의 wavelet을 적용하여 수행한 연구 결과, 본 시스템에서는 Haar 또는 Daubechies 웨이블릿을 사용하는 것이 가장 우수한 잡음소거 성능을 달성하는 것으로 나타났다. 아울러 MSE가 다른 웨이블릿을 사용하는 경우와 큰 차이가 있음을 보여주었다.

본 시스템과 관련한 향후 연구에서는 딥러닝의 최적화 알고리즘에 대한 성능 비교를 수행할 계획이다.

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] 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.

[3] H. Lee, "Nonlinear noise attenuator by adaptive Wiener filter with neural network," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 18, no. 1, 2023, pp. 71-76.

[4] J. 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.

[5] W. A. Harrison, J. 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.

[6] V. Justin, S. Saudia, and T. Nasser, "Fourier transform-based windowed adaptive switching minimum filter for reducing periodic noise from digital images," *IET image processing*, vol. 10, no. 9, 2016, pp. 646-656.

[7] I. Daubechies, "The Wavelet Transform Time-Frequency Localization and Signal Analysis," *IEEE Trans. on Information Theory*, vol. 36, no. 5, 1990, pp. 961-1005.

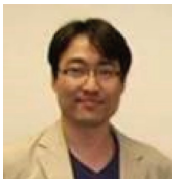
[8] C. Lee and D. Kim, "Adaptive Noise Reduction of Speech Using Wavelet Transform," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 4, no. 3, 2009, pp. 190-196.

[9] D. L. Dondio, "De-Noising by Soft-Thresholding," *IEEE Trans. on Information Theory*, vol. 41, no. 3, 1995, pp. 613-627.

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

[11] H. Lee, "Optimization of the number of filter in CNN noise attenuator," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 16, no. 4, 2021, pp. 625-632.

저자 소개



정원석(Won-Seok Jeong)

1991년 세명대학교 전자공학과(공학사)
2023년 수원대학교 글로벌창업대학원 창업경영학과 재학
2006년~2017년 ㈜테크엘 IoT사업본부 수석연구원
2017년~2019년 ㈜비엔컴 신사업 총괄 이사
2021년~현재 ㈜AICube 대표이사
※관심분야 : IT 모듈제작, 키오스크, 배경잡음 제거

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 지능정보통신공학과 교수
※ 관심분야 : VLSI 설계, 딥러닝, 음향잡음 소거