특집논문 (Special Paper)
방송공학회논문지 제30권 제6호, 2025년 11월 (JBE Vol.30, No.6, November 2025)
https://doi.org/10.5909/JBE.2025.30.6.918
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

뉴럴 보코더를 이용한 스펙트럼 크기 영역에서의 저지연 음성 향상 기술

유 현 석[a)], 유 정 찬[a)], 마 효 승[a)], 박 호 종[a)‡]

Low-delay Speech Enhancement using Neural Vocoder in Spectral Magnitude Domain

Hyunseok Yu[a)], Jeongchan Yu[a)], Hyoseung Ma[a)], and Hochong Park[a)‡]

요 약

본 논문에서는 뉴럴 보코더를 사용하여 스펙트럼 크기 영역에서 저지연으로 음성을 향상시키는 방법을 제안한다. 뉴럴 보코더는 HiFi-GAN을 기반으로 개발하였고, 지연 시간 조정을 위한 입력 계층을 추가하고 저지연 동작을 위한 세부 규격을 설계하였다. 제안 방법은 스펙트럼 크기를 시간 영역 파형으로 복원할 때 위상을 사용하지 않으므로 잘못된 위상에 의한 왜곡 문제를 해결할 수 있다. 또한, 제안 방법은 시간 영역 복원 과정에서 프레임 간 중첩이 필요하지 않으므로 동일한 프레임 길이 조건에서 기존 방법보다 저지연으로 작동할 수 있다. 주관적 청취 평가를 통하여 음성 향상 성능을 측정하였고, 기존 방법 대비 50%의 동작 지연을 가지는 제안 방법이 기존 방법보다 향상된 성능을 가지는 것을 확인하였다.

Abstract

In this paper, we propose a low-delay speech enhancement method using neural vocoder in the spectral magnitude domain. We developed the neural vocoder based on HiFi-GAN, and added a new input layer for delay control and designed new architecture for low-delay operation. When reconstructing the spectral magnitude into a time-domain waveform, the proposed method does not use phase information, thereby eliminating the degradation caused by incorrect phase. In addition, the proposed method can reduce the algorithmic delay, compared to the conventional method using the same frame length, because it does not use overlap process between frames. A subjective performance evaluation confirmed that the proposed method provides improved performance with 50% latency, compared to the conventional method.

Keyword : Speech enhancement, Neural vocoder, Low delay, Machine learning, GAN

a) 광운대학교 전자공학과(Dept. of Electronics Eng., Kwangwoon Univ.)
‡ Corresponding Author : 박호종(Hochong Park)
E-mail: hcpark@kw.ac.kr
Tel: +82-2-940-5104
ORCID: https://orcid.org/0000-0003-1600-6610
※이 논문의 결과 중 일부는 한국방송·미디어공학회 2025년 하계학술대회에서 발표한 바 있음
※이 논문은 2025년도 광운대학교 교내학술연구비 지원과 2025년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원(No. RS-2021-KI002499, 2025년 산업혁신인재성장지원사업)을 받아 수행된 연구임
· Manuscript September 11, 2025; Revised October 22, 2025; Accepted October 23, 2025.

Ⅰ. 서 론

음성 향상은 잡음 환경에서 음성 신호의 명료도를 향상하는 기술이다. 최근에는 머신러닝 기술의 발전으로 학습 기반의 음성 향상 연구가 활발히 이루어지고 있으며, 우수한 음성 향상 성능을 보이고 있다[1,2]. 머신러닝을 이용한 대표적인 음성 향상 기법인 스펙트럼 크기 영역에서의 방법은 잡음이 섞인 신호의 스펙트럼 크기에서 잡음 성분을 제거하고 잡음이 섞인 위상(noisy phase, NP)과 결합하여 시간 영역 파형으로 복원한다[3]. 이 방법에서 스펙트럼 크기와 위상은 각각 독립적으로 결정되고 서로 다른 신호의 크기 및 위상 성분에 해당하기 때문에, 시간 영역 복원 과정에서 왜곡이 발생한다.

NP로 인한 왜곡을 해결하기 위해 위상을 고려한 음성 향상 기술이 활발히 연구되고 있고, 복소 스펙트럼 영역, 시간 영역, 스펙트럼과 시간의 복합 영역에서의 음성 향상 방법이 개발되었다[4-6]. 이 방법들은 NP로 인한 문제를 해결하여 향상된 음성 향상을 제공할 수 있지만 스펙트럼 크기 영역 방법 대비 다음의 문제점을 가진다. 복소 스펙트럼 영역 방법은 복소 영역에서 잡음을 제거하므로 NP로 인한 문제를 근본적으로 해결할 수 있지만, 크기와 위상 사이의 복잡한 관계로 인하여 손실 함수를 정밀하게 설계하지 않으면 스펙트럼 크기의 왜곡을 일으키는 보상 효과(compensation effect) 문제를 유발한다[7]. 시간 영역 방법은 시간 영역에서의 낮은 희소도(sparsity)로 인해 잡음 분리를 위한 고난도 모델 설계를 요하고, 학습에 사용하지 않은 데이터에 대한 일반화가 힘들다는 단점을 가지며, 일반적으로 복소 스펙트럼 영역 방법에 비해 낮은 성능을 가진다[8]. 복합 영역 방법은 복소 스펙트럼과 시간 영역 모두에서 음성 향상을 수행하여 각 영역의 단점을 보완할 수 있지만, 두 영역의 결합을 위해 복잡한 모델 구조가 필요하고 인과적(causal)인 모델 설계에는 아직 한계를 가진다.

많은 음성 향상 방법은 비인과적(non-causal)으로 동작하거나 매우 긴 지연 시간을 가진다. 그러나 실시간 음성 통신을 위한 저지연 음성 향상의 수요가 늘어감에 따라 스펙트럼 기반의 음성 향상에서 동작 지연(algorithmic delay)을 줄이는 연구가 필요하다. 스펙트럼 기반 음성 향상에서는 스펙트럼을 시간 영역으로 복원하는 단구간 푸리에 역변환(inverse short-time Fourier transform, ISTFT)이 수행되고, 이 과정에서 프레임 중첩으로 인해 윈도우(window) 길이만큼의 동작 지연이 발생한다. 이를 해결하기 위해 미래 프레임을 예측하여 현재 프레임과 중첩하는 등의 저지연 음성 향상 연구 또한 진행되고 있다[9-11].

본 논문에서는 스펙트럼 크기 영역에서의 음성 향상에서 NP 문제와 지연 문제를 동시에 해결하기 위하여 뉴럴 보코더(neural vocoder)를 사용하는 음성 향상 방법을 제안한다. 뉴럴 보코더는 신경망을 통하여 스펙트럼 크기를 시간 영역으로 변환시키는 음성 합성 기술이고, 위상을 사용하지 않고 프레임 중첩이 필요 없으므로 NP 문제 해결과 저지연 동작이 가능하다. 제안 방법은 HiFi-GAN 기반의 뉴럴 보코더를 사용한다[12]. HiFi-GAN은 많은 음성 응용 분야에서 이용되는 대표적인 뉴럴 보코더이며, 우수한 음성 합성 성능을 제공하지만 수용 영역(receptive field)이 넓어 지연 시간이 길다는 문제를 가진다. 따라서 본 논문에서는 저지연 동작을 하도록 HiFi-GAN 동작을 수정하고 주어진 스펙트럼 규격에 맞추어 최종 뉴럴 보코더를 개발한다.

음성 향상의 성능 평가는 NSDTSEA 데이터셋을 사용하여 진행하였다[13]. 스펙트로그램 분석을 통하여 기존 방법에서 NP로 인해 발생하는 왜곡 문제를 제안 방법이 해결하는 것을 확인하였고, 객관적 성능 지표인 deep noise suppression mean opinion score (DNSMOS)를 사용하여 다양한 동작 지연 시간에 대하여 제안 방법이 기존 방법보다 우수한 성능을 가지는 것을 검증하였다[14]. 또한, 주관적 청취 품질 지표인 degradation mean opinion score (DMOS)를 사용하여 음성 향상 성능을 측정하였고[15], 기존 방법 대비 50% 동작 지연을 가지는 제안 방법이 기존 방법보다 청취적으로 우수한 품질의 음성을 제공하는 것을 확인하였다. 마지막으로, 제안 방법이 i7-10700 CPU와 RTX 3070 (8GB) GPU 환경에서 실시간으로 작동할 수 있음을 검증하였다.

Ⅱ. 제안하는 음성 향상 방법

1. 개발 배경

스펙트럼 크기 영역에서의 음성 향상 기술은 음성 신호

의 스펙트럼 크기에서 잡음 성분을 제거하여 향상된 스펙트럼 크기를 얻는 기술이다. 그림 1 (a)는 기존의 일반적인 스펙트럼 크기 영역 음성 향상 과정을 보여준다. 잡음이 섞인 시간 영역 신호에 단구간 푸리에 변환(short-time Fourier transform, STFT)을 적용하여 잡음이 섞인 스펙트럼 크기와 위상을 얻고, 스펙트럼 크기 영역 음성 향상 모델을 통해 잡음이 제거된 스펙트럼 크기를 얻고, 여기에 NP를 결합하고 ISTFT를 적용하여 시간 영역의 향상된 음성 신호로 복원한다. 이때, 별도로 처리하지 않은 NP를 시간 영역 복원 과정에서 그대로 사용하기 때문에, 스펙트럼 크기에서 완벽하게 잡음을 제거하여도 시간 영역으로 복원된 신호에는 청각적으로 인지 가능한 왜곡이 발생한다.

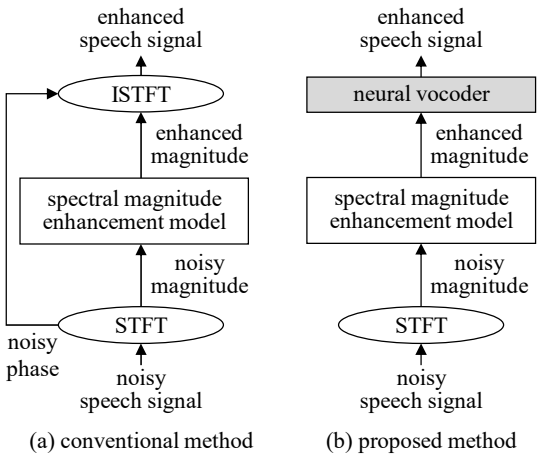

그림 1. 스펙트럼 크기 영역에서의 음성 향상 과정 (a) 기존 방법 (b) 제안 방법

Fig. 1. Speech enhancement in spectral magnitude domain (a) conventional method (b) proposed method

본 논문에서는 기존 방법에서 NP로 인해 발생하는 문제를 해결하기 위하여 그림 1 (b)와 같이 시간 영역 복원 과정에 뉴럴 보코더를 사용하는 방법을 제안한다. 뉴럴 보코더는 출력 신호가 프레임 사이에 연속성을 유지하도록 학습되기 때문에, 직접적으로 위상을 사용하지 않고 스펙트럼 크기에 맞는 위상을 자동으로 적용하는 것과 같이 동작한다. 따라서 제안 방법은 향상된 스펙트럼 크기를 시간 영역 파형으로 복원하는 과정에서 NP로 인한 왜곡 문제를 해결할 수 있다.

또한, 제안 방법은 기존 방법보다 동작 지연을 줄일 수 있다. 총 동작 지연은 현재 프레임 길이에 해당하는 프레임 지연과 연산 과정에서 필요한 미래 시점의 프레임 길이에 해당하는 미리 보기(look-ahead) 지연으로 이루어진다. 기존 방법은 스펙트럼 크기의 시간 영역 복원 과정에서 프레임 중첩이 필요하고 미래 시점의 프레임을 사용하므로 미리 보기 지연이 필요하며, 이론적으로 윈도우 길이만큼의 동작 지연을 가진다. 반면, 제안 방법의 뉴럴 보코더는 복원 과정에서 프레임 중첩이 필요하지 않으며, 미리 보기 지연을 자유롭게 설계할 수 있다. 따라서 본 논문에서는 기존 방법의 긴 동작 지연 문제를 해결하기 위해 저지연 뉴럴 보코더를 설계하여 음성 향상에 적용한다.

제안 방법은 특정 모델 구조에 한정되지 않으며, 그림 1 (b)와 같이 임의의 스펙트럼 크기 영역 음성 향상 모델과 임의의 뉴럴 보코더를 결합하여 구현할 수 있다. 본 논문에서는 간단한 스펙트럼 크기 영역 음성 향상 모델을 사용하여 제안 방법이 기존 방법의 NP로 인한 문제와 동작 지연 문제를 해결하고 기존 방법 대비 우수한 음성 향상 성능을 가지는지 검증한다.

2. 스펙트럼 크기 영역 음성 향상을 위한 신경망

본 논문에서는 기존 방법과 제안 방법을 비교하기 위하여 두 방법에 동일한 구조의 스펙트럼 크기 영역 음성 향상 모델을 사용한다. 그림 2는 검증에 사용한 스펙트럼 크기

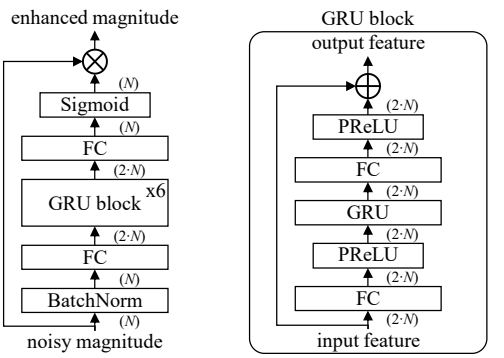

그림 2. 성능 평가를 위해 사용한 스펙트럼 크기 영역 음성 향상 모델 구조

Fig. 2. Structure of spectral magnitude enhancement model used for performance evaluation

영역 음성 향상 모델 구조를 보여준다. 게이트 순환 유닛(gated recurrent unit, GRU)과 완전 연결 계층(fully connected layer, FC) 기반의 마스킹(masking) 모델을 사용하고, N은 입력 스펙트럼 크기의 빈(bin) 개수이다. 모델은 프레임 단위로 동작하고, 잡음이 섞인 스펙트럼 크기에서 N개 빈의 스펙트럼 크기를 입력으로 하며, 잡음 성분이 제거된 N개 빈의 향상된 스펙트럼 크기를 인과적으로 출력한다.

3. 시간 영역 복원을 위한 저지연 뉴럴 보코더

음성 향상 모델의 출력인 스펙트럼 크기를 시간 영역 신호로 복원하기 위한 저지연 뉴럴 보코더는 음성 합성 기술에서 널리 사용되고 있는 HiFi-GAN을 활용하여 설계하였다[12]. 그림 3은 본 논문에서 사용한 뉴럴 보코더의 구조를 보여주며, 지연 시간 조절을 위한 입력 합성곱 계층(convolution layer)과 저지연 동작에 맞도록 변경된 HiFi-GAN 생성자(generator)로 구성된다. 이때, c와 k는 각각 합성곱 계층의 출력 채널(channel) 수와 커널 크기(kernel size)이고, a는 활성화 함수, N은 입력 스펙트럼 크기의 빈 개수, k_u는 업샘플링(upsampling)을 위한 전치 합성곱 계층(transposed convolution layer)의 커널 크기, l은 뉴럴 보코더의 미리 보기 지연 조절을 위한 매개 변수이다.

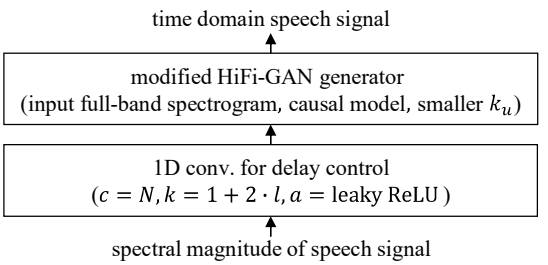

그림 3. 제안 방법의 저지연 뉴럴 보코더 구조
Fig. 3. Structure of low-delay neural vocoder in the proposed method

원 HiFi-GAN 생성자는 80개 대역으로 이루어진 멜(mel) 스펙트럼을 입력하여 시간 영역 신호를 생성하도록 설계되었으며 비인과적으로 동작한다. 본 논문에서는 뉴럴 보코더가 ISTFT와 같이 모든 대역의 스펙트럼 크기를 입력으로 하고 인과적으로 동작할 수 있도록 HiFi-GAN 생성자의 동작을 변경하였다. 먼저, 모든 스펙트럼 빈의 크기를 입력하기 위해 수정된 HiFi-GAN 생성자의 입력 채널 수를 N으로 조정하였다. 본 논문에서는 512-포인트 DFT (discrete Fourier transform)를 취하여 얻은 256개 빈의 스펙트럼 크기를 뉴럴 보코더의 입력으로 사용하므로, N을 256으로 설정하였다. 다음, 수정된 HiFi-GAN 생성자의 모든 합성곱 계층과 전치 합성곱 계층의 패딩(padding)을 과거 방향에만 적용하여 인과적인 동작을 구현하였다.

추가로, HiFi-GAN 생성자에서 k_u를 [16, 8, 4, 4]로 수정하였다. 원 HiFi-GAN에서는 한 프레임의 스펙트럼 크기에서 256 샘플(sample)을 생성하나, 본 논문에서는 한 프레임의 스펙트럼 크기에서 128 샘플을 생성한다. 따라서 업샘플링의 역할을 하는 전치 합성곱 계층 중 두 번째 전치 합성곱 계층의 보폭(stride)을 원 HiFi-GAN 생성자 대비 절반으로 설정하였다. 원 HiFi-GAN에서 사용된 전치 합성곱 계층의 커널 크기는 보폭의 두 배이므로 보폭을 수정한 계층의 커널 크기 또한 원 HiFi-GAN의 절반으로 설정하였다. 이외의 구조는 원 HiFi-GAN의 V1 생성자와 동일하게 설정하였다[12].

기존 방법과 제안 방법의 동등한 성능 비교를 위하여 두 방법은 스펙트럼 크기의 시간 영역 복원 과정에서 동일한 동작 지연이 필요하고, 뉴럴 보코더에 ISTFT와 동일한 미리 보기 지연이 필요하다. 그림 3과 같이 동작 지연 조절을 위한 합성곱 계층을 설계하여 수정된 HiFi-GAN 생성자 앞 단에 추가하였다. 해당 합성곱 계층은 한 단의 합성곱 계층으로 이루어져 있고, 커널 크기 조절을 통하여 한 프레임의 프레임 지연과 l 프레임만큼의 미리 보기 지연을 구현하였다.

뉴럴 보코더가 자연스러운 음성 신호를 생성할 수 있게 학습 과정에 판별자(discriminator)를 사용하였다. 판별자 구성은 BigVGAN과 동일하게 HiFi-GAN의 MPD (multi-period discriminator)와 UnivNet의 MRSD (multi-resolution spectrogram discriminator)를 사용하였다[12,16,17].

Ⅲ. 성능 평가

1. 데이터셋과 학습 방법

모든 신경망 학습과 성능 평가는 NSDTSEA 데이터셋을

사용하여 진행하였다[13]. NSDTSEA 학습 데이터셋은 DEMAND 데이터셋에서 추출한 8가지 환경 잡음과 2가지 인공 잡음을 VoiceBank corpus에서 추출한 화자 28명의 발화 데이터와 15dB, 10dB, 5dB, 0dB의 SNR (signal noise ratio)로 혼합하여 만든 11,572개, 총 9시간 20분가량의 데이터로 구성되어 있다. 평가 데이터셋은 학습 데이터와 중복되지 않는 VoiceBank corpus에서 추출한 화자 2명의 발화 데이터와 DEMAND 데이터셋에서 추출한 5가지 환경 잡음 데이터를 17.5dB, 12.5dB, 7.5dB, 2.5dB로 혼합하여 만든 824개, 총 30분가량의 데이터로 구성되어 있다.

본 논문에서는 각 데이터를 16kHz로 다운샘플링(down-sampling)하여 학습 및 평가에 사용하였다. 학습 과정에서 모델의 입력 데이터는 각 파일마다 매 에포크(epoch)에서 무작위로 16,384 샘플(1.024초) 길이의 구간을 추출하여 STFT를 취한 스펙트럼 크기를 사용하였다. 이때, 프레임은 128 샘플, STFT는 512 샘플, 사인 윈도우(sine window), 512-포인트 DFT, 75% 중첩 규격을 적용하였다.

스펙트럼 크기 영역 음성 향상 모델과 뉴럴 보코더는 독립적으로 학습된다. 학습 과정에서 스펙트럼 크기 영역 음성 향상 모델은 잡음이 섞인 음성 신호를 입력하고, 뉴럴 보코더는 깨끗한 음성 신호를 입력한다. 두 모델 모두 깨끗한 음성 신호를 목표(target) 데이터로 사용한다.

제안 방법은 그림 1 (b)와 같이 두 모델이 순차적으로 작동하도록 결합하여 구현된다. 따라서 추론 단계에서 뉴럴 보코더 입력은 음성 향상 모델이 제공한 향상된 스펙트럼 크기이며, 이로 인하여 뉴럴 보코더의 학습-추론 사이에 불합치 문제가 발생한다. 이를 해결하기 위하여 결합 모델에 대하여 미세 학습(fine tuning)을 진행한다. 미세 학습 과정에서 각각 학습된 두 모델을 결합하여 종단간(end-to-end) 학습하고, 결합 모델의 입력은 잡음이 섞인 음성 신호이고, 목표 데이터는 깨끗한 음성 신호이다. 이를 통해 음성 향상 모델은 뉴럴 보코더에 최적화된 향상된 스펙트럼 크기를 출력하도록 학습하고, 뉴럴 보코더는 향상된 스펙트럼 크기에 대하여 최적 동작을 하도록 학습한다. 또한, 미세 학습 과정에서 판별자는 뉴럴 보코더의 독립적인 학습 과정에서 학습된 판별자가 아닌, 초기 상태의 판별자를 사용한다.

음성 향상 모델의 학습은 0.0003 초기 학습률(learning rate)에 0.98배의 학습률 감쇠(learning rate decay)를 매 에포크마다 적용하여 300 에포크 동안 진행하였고, 뉴럴 보코더의 학습은 0.0002 초기 학습률에 0.99배의 학습률 감쇠를 10 에포크마다 적용하여 500 에포크 동안 진행하였다. 미세 학습에서는 0.00005 초기 학습률에 0.99배의 학습률 감쇠를 10 에포크마다 적용하여 200 에포크 동안 학습하였다. 모든 학습 과정에서 최적화기(optimizer)는 AdamW에 $\beta_1 = 0.8, \beta_2 = 0.99$를 적용하여 사용하였고, 미니 배치(mini-batch) 크기는 16으로 설정하였다. 음성 향상 모델 학습에서의 손실 함수는 모델 출력과 목표 데이터의 스펙트럼 크기 간 평균 절대 오차이고, 뉴럴 보코더 학습에서의 손실 함수는 원 HiFi-GAN의 손실 함수에서 멜 스펙트럼과 관련된 항을 스펙트럼 빈의 크기에 대한 항으로 수정하여 활용하였다[12]. 결합 모델의 미세 학습은 뉴럴 보코더와 동일한 손실 함수를 사용하였고, 결합 모델을 하나의 생성 모델로 취급하여 손실 함수를 적용하였다.

2. 성능 검증

음성 향상의 성능 비교를 위한 기존 모델은 그림 2의 신경망을 그림 1 (a)와 같이 활용하여 동작하며, 제안 방법은 N을 256으로 설정한 그림 2 신경망과 그림 3의 뉴럴 보코더를 그림 1 (b)와 같이 결합하여 동작하고, 기존 모델을 Base, 제안 모델을 Prop이라 표기한다. 다양한 동작 지연 시간에 대한 성능을 평가하고, 동일한 동작 지연에 대하여 기존 모델과 제안 모델의 성능을 비교하기 위하여 3가지 동작 지연을 가지도록 각 모델의 세부 동작을 설계하였다. 기존 모델에서 프레임 길이가 64, 96, 128이고, 75% 윈도우 중첩을 사용하고, DFT 포인트를 256, 384, 512으로 설정하고, 그림 2의 N을 128, 192, 256으로 설정하여 각각 16ms, 24ms, 32ms의 동작 지연을 가지는 Base 모델을 구현하였고, 각각 Base16, Base24, Base32로 표기한다.

제안 방법에 사용하는 뉴럴 보코더는 동작 지연 조절이 자유로운 구조를 가지며, 그림 3에서 동작 지연 조절을 위한 합성곱 계층의 l을 1, 2, 3으로 설정하여 각각 16ms, 24ms, 32ms의 동작 지연을 가지는 제안 모델을 구현하였다. 또한, 기존 방법과 제안 방법의 주요 차이점인 ISTFT와 뉴럴 보코더 사이의 직접적 성능 비교를 위해 동일한 향상된 스펙트럼 크기를 활용하는 모델 간의 비교가 필요하다.

따라서 미세 학습 없이, 독립적으로 학습된 음성 향상 모델과 뉴럴 보코더를 단순 결합한 제안 모델을 검증 과정에 사용하였고, 이 모델을 각 동작 지연에 따라 Prop16, Prop24, Prop32로 표기한다. 마지막으로, 성능 향상을 위해 미세 학습을 진행한 제안 모델은 각 동작 지연에 따라 Prop16*, Prop24*, Prop32*로 표기한다.

제안 방법의 뉴럴 보코더로 사용한 HiFi-GAN은 생성형 모델이기 때문에 목표 신호와 정확하게 일치하는 신호 출력을 목표로 하지 않는다. 이로 인하여 출력 신호와 목표 신호를 비교하는 것은 의미가 없고, 생성된 신호가 목표 신호의 특징을 유지하고 청취적으로 고품질을 가지도록 하는 것이 목표이다. 따라서 제안 방법의 성능 평가를 위해 PESQ와 같이 원본 신호와 평가 신호를 비교하는 방식의 객관적 평가 지표는 사용할 수 없으며, 원본 신호 없이 평가 신호의 청취 품질을 측정하는 비침입적(non-intrusive) 평가 지표인 DNSMOS와 DMOS를 사용하였다[14,15,18,19].

2.1 객관적 성능 평가

스펙트로그램 분석을 통해 제안 방법이 기존 방법의 NP 문제를 해결할 수 있음을 확인하였다. 그림 4는 서로 다른 신호에 대하여 왼쪽부터 각각 잡음이 섞인 신호, 향상된 스펙트럼 크기, Base32와 Prop32를 통하여 잡음을 제거한 신호, 잡음이 없는 깨끗한 신호의 스펙트로그램을 보여준다. 이때, 향상된 스펙트럼 크기는 음성 향상 모델이 출력한 스펙트럼 크기를 시각화한 것이고, 나머지 항목은 최종 출력된 시간 영역 신호로부터 스펙트로그램을 구하여 나타낸 것이다.

그림 4 (a)의 잡음이 포함된 신호는 0.6 – 1.2kHz 대역에 톤 성질(tonal)의 잡음이 섞여 있는 신호이고, 향상된 스펙트럼 크기에서 음성 향상 모델을 통하여 잡음 성분이 지워진 것을 확인할 수 있다. 그러나 Base32에서 NP로 인해 출력 신호에 잡음 성분과 유사한 신호가 복원된 것을 확인할 수 있다. 반면에 Prop32에서 뉴럴 보코더가 향상된 스펙트럼 크기에 대응하는 음성 신호를 생성하기 때문에 Base32의 결과에 나타나는 왜곡이 발생하지 않는다. 또한, 그림 4 (b)의 향상된 스펙트럼 크기의 0.6 – 1.8kHz 대역에는 하모닉 성분이 있는 것을 확인할 수 있으나, Base32에서 NP로 인한 하모닉 성분의 왜곡을 확인할 수 있다. 반면에 Prop32에서는 향상된 스펙트럼 크기의 하모닉 성분이 복원된 것을 확인할 수 있다.

표 1은 모든 음성 향상 모델의 DNSMOS 성능을 보여준다. 각 동작 지연에 대하여 제안 모델이 기존 모델보다 DNSMOS 성능이 높은 것을 확인할 수 있고, 특히 제안 모델 중 가장 짧은 동작 지연을 가지는 Prop16와 Prop16*이

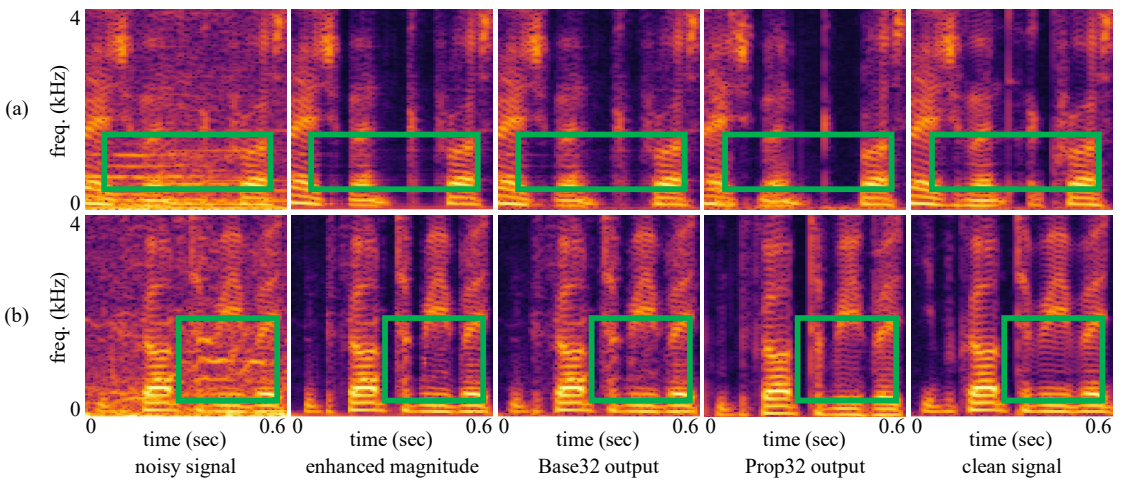

그림 4. 잡음이 포함된 신호, 향상된 스펙트럼 크기, Base32와 Prop32의 출력 신호, 깨끗한 신호의 스펙트로그램 예시, (a) 톤 성질의 잡음이 포함된 신호의 예시 (b) 다화자 잡음이 포함된 신호의 예시
Fig. 4. Examples of noisy signal, enhanced magnitude, output of Base32 and Prop32, and clean signal, (a) examples for signals containing tonal noise (b) examples for signals containing babble noise

기존 모델 중 가장 긴 동작 지연을 가지는 Base32보다 더 높은 DNSMOS 성능을 가짐을 확인할 수 있다. 또한, 미세 학습을 진행한 모델이 미세 학습을 적용하지 않은 모델에 비하여 성능이 향상된 것을 확인할 수 있고, 이 결과로부터 미세 학습을 통해 뉴럴 보코더 학습에서 발생하는 학습-추론 불합치 문제가 완화됨을 알 수 있다. 모델의 부동 소수점 연산 횟수(floating point operations, FLOPs)는 fvcore 라이브러리를 사용하여 1초 분량의 신호를 처리하기 위해 필요한 연산 횟수를 측정하였다[20].

표 1. 각 음성 향상 모델의 DNSMOS 결과
Table 1. DNSMOS results for different speech enhancement models

algorithmic delay	model	DNSMOS	# of parameters	FLOPs
	clean signal	3.55		
	noisy signal	3.03		
16ms	Base16	3.30	3.23M	0.21G
	Prop16	3.37	27.37M	24.00G
	Prop16*	3.47	27.37M	24.00G
24ms	Base24	3.31	7.25M	0.32G
	Prop24	3.43	27.50M	24.01G
	Prop24*	3.48	27.50M	24.01G
32ms	Base32	3.32	12.88M	0.43G
	Prop32	3.44	27.63M	24.03G
	Prop32*	3.50	27.63M	24.03G

제안 방법에서 실시간 작동이 가능하도록 복잡도를 감소시킨 Prop16*(small)을 추가로 구현하였다. 그림 3의 뉴럴 보코더 구조에서 N, k_u, l은 각각 256, [16, 16, 4], 1로 설정하였고, 수정된 HiFi-GAN 생성자의 첫 합성곱 계층의 출력 채널을 512로 설정하였다. 그 외의 설정은 원 HiFi-GAN의 V3 생성자와 동일하게 구성하였다[12].

Prop16*(small)이 실시간으로 작동하는지 검증하기 위하여 i7-10700 CPU와 RTX 3070 (8GB) GPU 환경에서 입력 샘플 길이 대비 작동 시간인 RTF (real-time factor)를 측정하였다. 해당 모델의 RTF는 0.91로 1 미만의 RTF를 달성하여 실시간 작동이 가능함을 검증하였다. Prop16*(small)의 DNSMOS는 3.45이며, 기존 방법 중 가장 동작 지연이 긴 Base32보다 더 높은 DNSMOS 성능을 가짐을 확인할 수 있다. Prop16*(small)의 모델 파라미터 수는 21.81M이고, 연산량은 17.36 GFLOPs이다.

2.2 주관적 성능 평가

주관적 성능 평가는 DMOS를 사용하여 진행하였고, 청취자가 표 2의 기준에 따라 원본 신호 대비 평가 신호가 얼마나 왜곡되었는지를 평가한다[15]. 평가에는 NSDTSEA 평가 데이터셋에서 각 SNR별로 5개를 무작위 추출하여 구성한 총 20개의 잡음이 섞인 음성 신호와 각각에 대한 Base32, Prop16*, Prop16*(small)의 출력 신호를 사용하였다. 7명의 훈련된 평가자가 성능 평가에 참여하였으며, 반복 없이 1초 간격으로 잡음이 섞이지 않은 원본 신호와 평가 신호를 듣고 평가 신호의 왜곡 점수를 부여하였다.

표 2. DMOS 평가 기준
Table 2. Scoring criteria of DMOS

score	scoring criteria
5	Degradation is inaudible.
4	Degradation is audible but not annoying.
3	Degradation is slightly annoying.
2	Degradation is annoying.
1	Degradation is very annoying.

표 3은 DMOS 결과를 정리한 것이고, Prop16*과 Prop16*(small)은 모두 두 배의 동작 지연 시간을 가지는 Base32보다 높은 DMOS 성능을 가진다. 이를 통해 기존 방법 대비 50% 동작 지연을 가지는 제안 방법이 기존 방법에 비하여 청취적으로 우수한 음성 향상 성능을 가지고, 실시간 작동이 가능함을 검증하였다.

표 3. 기존 방법과 제안 방법의 DMOS 결과
Table 3. DMOS results of conventional method and proposed method

model	DMOS	# of parameters	FLOPs
noisy signal	1.61		
Base32	2.44	12.88M	0.43G
Prop16*	3.50	27.37M	24.00G
Prop16*(small)	3.04	21.81M	17.36G

Ⅳ. 결 론

본 논문에서는 기존 음성 향상 기술의 위상으로 인한 왜곡과 지연 문제를 해결하기 위해 뉴럴 보코더를 이용한 스

펙트럼 영역에서의 저지연 음성 향상 방법을 개발하였다. 제안 방법은 향상된 스펙트럼 크기의 시간 영역 복원 과정에서 별도의 위상을 사용하지 않고, 프레임 중첩이 필요하지 않아 기존 방법의 위상으로 인한 왜곡과 지연 문제를 동시에 해결할 수 있음을 확인하였다. 기존 방법 대비 50%의 동작 지연을 가지는 음성 향상 모델을 구현하였고, 주관적 청취 평가를 통해 제안 방법이 더 높은 음성 향상 성능을 가짐을 확인하였다. 또한, 제안 방법이 i7-10700 CPU와 RTX 3070 (8GB) GPU 환경에서 실시간 작동이 가능함을 검증하였다.

향후 동작 지연 시간을 더 줄이기 위한 DFT 구조 변경과 미리 보기 미사용 등의 연구와 이때 발생하는 주파수 해상도 저하, 연산량 증가, 시스템 안정성 감소 등에 대한 상호 보완적인 연구가 필요하다. 또한 본 논문에서 검증한 CPU와 GPU 환경에서의 실시간 작동 가능성을 바탕으로 이동 통신 환경과 같이 제한적인 환경에서 실시간으로 작동하는 모델 구조 연구가 필요하다. 이러한 방향성이 제안 방법의 실제 적용 범위를 확장하는 데 핵심적인 역할을 할 것으로 기대된다.

참 고 문 헌 (References)

[1] J. Yu, J. Kim, H. Moon and H. Park, “Speech enhancement based on speech production model using machine learning,” *J. of Broadcast Engineering*, Vol. 28, No. 6, pp. 743-752, 2023.
doi: https://doi.org/10.5909/JBE.2023.28.6.743

[2] S. Hwang, J. Byun, J. Heo, J. Cha, and Y. Park, “Multi-level skip connection for nested U-Net-based speech enhancement,” *J. of Broadcast Engineering*, Vol. 27, No. 6, pp. 840-847, 2022.
doi: https://doi.org/10.5909/JBE.2022.27.6.840

[3] C. Zheng et al., “Sixty years of frequency-domain monaural speech enhancement: from traditional to deep learning methods,” *Trends Hearing*, vol. 27, No. 23312165231209913, 2023.
doi: https://doi.org/10.1177/23312165231209913

[4] Ye-Xin Lu, Ai Yang and Zhen-Hua Ling, “MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra,” *in Proc. Interspeech*, Dublin, Ireland, pp. 3834-3838, 2023.
doi: https://doi.org/10.21437/Interspeech.2023-1441

[5] A. Pandey and D. Wang, “TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain,” *in Proc. ICASSP*, Brighton, UK, pp. 6875-6879, 2019.
doi: https://doi.org/10.1109/ICASSP.2019.8683634

[6] J. Kim, J. Yoo, S. Chun. A. Kim and J. Ha, “Multi-domain processing via hybrid denoising networks for speech enhancement,” *arXiv preprint*, arXiv:1812.08914, 2018.
doi: https://doi.org/10.48550/arXiv.1812.08914

[7] Z. Wang, G. Wichern and J. Le Roux, “On the compensation between magnitude and phase in speech separation,” *IEEE Signal Processing Letters*, vol. 28, pp. 2018‑2022, 2021.
doi: https://doi.org/10.1109/LSP.2021.3116502

[8] S. Nossier, J. Wall, M. Moniri, C. Glackin and N. Cannings, “A comparative study of time and frequency domain approaches to deep learning based speech enhancement,” *in Proc. IJCNN*, Glasgow, UK, pp. 1-8, 2020.
doi: https://doi.org/10.1109/IJCNN48605.2020.9206928

[9] Z. Wang, G. Wichern, S. Watanabe and J. Le Roux, “STFT-domain neural speech enhancement with very low algorithmic latency,” *in IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 397-410, 2023.
doi: https://doi.org/10.1109/TASLP.2022.3224285

[10] A. Dementyev et al., “Towards sub-millisecond latency real-time speech enhancement models on hearables,” *in Proc. ICASSP*, Hyderabad, India, 2025.
doi: https://doi.org/10.1109/ICASSP49660.2025.10889875

[11] H. Bae et al., “Speech boosting: Low-latency live speech enhancement for TWS earbuds,” *in Proc. Interspeech*, Kos, Greece, pp.647-651, 2024.
doi: https://doi.org/10.21437/Interspeech.2024-1444

[12] J. Kong, J. Kim and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” *in Advances in NIPS*, Vol.33, pp. 17022-17033, 2020.

[13] C. Valentini-Botinhao, “Noisy speech database for training speech enhancement algorithms and TTS models,” University of Edinburgh, School of Informatics, Centre for Speech Technology Research (CSTR), 2016.
doi: https://doi.org/https://doi.org/10.7488/ds/2117

[14] C. K. A. Reddy, V. Gopal and R. Cutler, “Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” *in Proc. ICASSP*, Toronto, Canada, pp. 6493-6497, 2021.
doi: https://doi.org/10.1109/ICASSP39728.2021.9414878

[15] ITU-T, *Methods for subjective determination of transmission quality*, P.800, 1996.

[16] S. Lee et al., “BigVGAN: A universal neural vocoder with large-scale training,” *in Proc. ICLR*, 2023.

[17] W. Jang et al., “UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation,” *in Proc. Interspeech*, Brno, Czechia, pp.2207-2211, 2021.
doi: https://doi.org/10.21437/Interspeech.2021-1016

[18] ITU-T, *Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, P.862, 2001.

[19] DNSMOS github, https://github.com/microsoft/DNS-Challenge/tree/master/DNSMOS (accessed July 23, 2025)

[20] fvcore github, https://github.com/facebookresearch/fvcore (accessed Oct. 18, 2025)

저 자 소 개

유 현 석

- 2024년 2월 : 광운대학교 전자공학과 공학사
- 2024년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : https://orcid.org/0009-0006-3397-257X
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

유 정 찬

- 2021년 2월 : 광운대학교 전자공학과 공학사
- 2023년 2월 : 광운대학교 전자공학과 공학석사
- 2023년 3월 ~ 현재 : 광운대학교 전자공학과 박사과정
- ORCID : https://orcid.org/0000-0003-0441-1280
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

마 효 승

- 2024년 2월 : 광운대학교 전자공학과 공학사
- 2024년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : https://orcid.org/0009-0009-8614-389X
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

박 호 종

- 1986년 2월 : 서울대학교 전자공학과 공학사
- 1987년 12월 : Univ. of Wisconsin-Madison 공학석사
- 1993년 5월 : Univ. of Wisconsin-Madison 공학박사
- 1993년 9월 ~ 1997년 8월 : 삼성전자 선임연구원
- 1997년 9월 ~ 현재 : 광운대학교 전자공학과 교수
- ORCID : https://orcid.org/0000-0003-1600-6610
- 주관심분야 : 오디오/음성 신호처리, 3D 오디오, 음악정보처리