특집논문 (Special Paper)
방송공학회논문지 제28권 제6호, 2023년 11월 (JBE Vol.28, No.6, November 2023)
https://doi.org/10.5909/JBE.2023.28.6.743
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

머신러닝을 이용한 음성 생성 모델 기반의 음성 향상 기술

유 정 찬[a)], 김 재 원[a)], 문 희 연[a)], 박 호 종[a)‡]

Speech Enhancement based on Speech Production Model using Machine Learning

Jeongchan Yu[a)], Jaewon Kim[a)], Heeyoun Moon[a)], and Hochong Park[a)‡]

요 약

본 논문은 음성 생성 모델에 따라 음성 향상을 수행하는 새로운 신경망 구조를 제안한다. 신경망은 입력신호로부터 여기신호와 스펙트럼 포락선을 구하고 각 성분에 대한 품질 향상을 수행하여 출력을 생성한다. 이 때, 각 성분의 특성에 맞는 제약조건을 신경망에 적용하여 음성 생성 모델에 따른 동작을 학습시킨다. 또한, 제안 방법은 음성에 특화된 제한적 동작을 수행하므로 기존 방법에 비해 신경망 복잡도를 감소시킨다. NSDTSEA 데이터셋을 사용하여 신경망 학습과 성능 평가를 진행하였고, 스펙트로그램 분석을 통하여 학습된 신경망이 음성 생성 모델에 따라 동작하여 음성 향상을 수행하는 것을 확인하였다. 또한 객관적 성능평가를 통해 제안 방법이 SEGAN과 WaveNet에 비해 각각 1,344배와 70배 적은 신경망 매개변수를 가지고 더 우수한 품질의 음성을 생성하는 것을 확인하였다. 이를 통해 제안 방법이 음성 생성 모델을 이용하여 적은 양의 신경망 매개변수로도 효율적인 음성 향상을 수행할 수 있음을 확인하였다.

Abstract

This paper proposes a new neural network architecture for speech enhancement based on speech production model. The network decomposes the input into the excitation signal and spectral envelope, and synthesizes the output after enhancing each component. Constraints appropriate for each component is applied to the network for the intended learning according to the speech production model. In addition, the proposed method conducts limited operations specific to speech, thus reducing the complexity compared with conventional methods. The NSDTSEA dataset is used for network training and performance evaluation, and the spectrogram analysis confirms that the learned network performs speech enhancement according to the speech production model. An objective performance evaluation confirms that the proposed method provides higher performance than the SEGAN and WaveNet, while using 1,344 and 70 times fewer network parameters than the SEGAN and WaveNet, respectively. These results verify that the proposed method can perform effective speech enhancement even using a small network owing to the speech production model.

Keyword : speech enhancement, noise suppression, speech production model, speech synthesis, interpretable machine learning

Ⅰ. 서 론

잡음이 포함된 음성신호에서 잡음을 제거하여 음성신호의 품질을 향상시키는 음성 향상 기술은 음성 통신 및 방송, 미디어 분야에서의 핵심적인 기술 중 하나로 오랜 기간 꾸준한 수요를 가져왔다[1,2]. 최근 머신러닝 기술이 발전함에 따라 전통적인 신호처리 기술에서 벗어나 학습을 통해 음성 향상 방법을 설계하는 연구가 널리 진행되고 있고, 일반적으로 머신러닝 기반의 음성 향상 기술이 신호처리 기반의 기술보다 우수한 성능을 보인다[3].

머신러닝을 이용한 음성 향상은 다양한 구조와 방법으로 시도되고 있다. 스펙트로그램(spectrogram) 영역에서 잡음 마스크(mask)를 예측하여 잡음 성분을 제거하는 방법[4], U-Net을 이용하여 신호를 다양한 해상도에서 분석하여 품질 향상된 음성신호를 생성하는 방법[5,6], 넓은 시간영역에서의 시간 의존성을 분석하는 WaveNet을 이용하여 고품질 음성신호를 생성하는 방법[7], 적대적 생성 신경망(generative adversarial network, GAN)을 이용하는 방법[8], 또는 이러한 방법들을 복합적으로 이용하는 방법 등이 있다[9,10]. 그러나 이들은 기존 머신러닝 분야에서 널리 사용되고 효과가 검증된 기법을 음성신호에 적용한 것이고, 음성신호에 특화된 최적의 방법이 아니다. 예로, WaveNet은 음성신호가 과거 신호에 의존성을 가지는 특징을 이용하지만 이는 음성신호만의 고유 특징을 고려한 것이 아니라 오디오 신호가 가지는 보편적인 특징을 이용한 것에 가깝다.

음성신호는 신호를 생성하는 고유의 생성 모델을 가지며, 이는 다른 신호들과 구분되는 뚜렷한 특징이다. 음성 생성 모델(speech production model)에 의하면, 음성신호는 성대(vocal cord)의 출력에 해당하는 여기신호(excitation signal)와 성도(vocal tract)의 공명 동작을 정의하는 스펙트럼 포락선(spectral envelope)의 곱으로 생성된다[11]. 유성음(voiced sound)은 성대를 진동시키며 생성되므로 유성음에 대한 여기신호는 음고(pitch)에 해당하는 기본 주파수(fundamental frequency)의 배음(harmonic)신호가 된다. 반면, 무성음(unvoiced sound)은 성대의 진동 없이 생성되고 이 때의 여기신호는 백색 잡음신호에 해당한다. 따라서 여기신호와 스펙트럼 포락선을 각각 고유의 성질에 맞게 생성하고 이를 결합하는 과정으로 음성신호를 생성할 수 있고, 이렇게 생성된 음성신호는 잡음 신호와 구분되는 고유한 특성을 가진다.

본 논문은 음성신호의 고유한 특징과 음성 생성 모델을 기반으로 음성신호에 특화된 잡음 제거 방법을 제안한다. 제안 방법은 신경망(neural network)을 사용하여 잡음이 제거된 여기신호와 스펙트럼 포락선을 각각 구하고 두 신호를 곱하여 잡음이 제거된 음성신호를 생성하며, 기존 머신러닝 방법에서 시도되지 않은 새로운 구조로 음성 향상을 수행한다. 또한, 제안 방법은 음성에 특화된 제한적 동작만을 수행하므로 기존 방법에 비해 신경망 복잡도를 감소시킨다. 즉, 본 논문의 목표는 음성 향상 성능을 높이는 새로운 머신러닝 기술을 개발하는 것이 아니라, 음성 생성 모델을 기반으로 음성 향상을 수행하는 새로운 방법론을 개발하고 동작의 효율성과 성능을 확인하는 것이다. 따라서 본 논문에서 제안하는 핵심 기술은 두 가지로 정리된다. 첫 번째는 음성 생성 모델에 따른 동작을 위해 두 성분의 곱 형태로 음성을 생성하는 신경망 구조를 설계하는 것이고, 두 번째는 각 성분의 특성에 따른 제약조건을 신경망에 적용하여 음성 생성 모델 기반의 동작을 정확히 수행하고 복잡도를 줄이는 것이다. 두 가지 기술을 통해 신경망은 음성 생성 모델에 따라 신호를 분할 및 합성하고 적은 양의 신경망 매개변수(parameter)로도 효율적으로 잡음 제거를 수행할 수 있다.

제안한 방법의 학습과 성능평가에는 NSDTSEA (noisy speech database for training speech enhancement algorithms and TTS models) 데이터셋을 사용하였고[12], 다양한 조건에서의 동작 비교를 위해 신경망 매개변수 개수와 제약조건 여부 등을 변경하여 신경망을 학습하고 성능을 평가하였다. 스펙트로그램 분석을 통해 신경망이 음성 생성 모델에 따라 동작하는지 검증하였고, PESQ (perceptual

a) 광운대학교 전자공학과(Dept. of Electronics Engineering, 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 논문은 2023년도 광운대학교 교내학술연구비 지원과 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2021R1F1A1059233)을 받아 수행된 연구임.
· Manuscript November 14, 2023; Revised November 22, 2023; Accepted November 22, 2023.

evaluation of speech quality), CSIG, CBAK, COVL를 사용하여 음성 향상 성능을 평가하였다[13]. 그 결과, 제안하는 신경망은 추가적인 제약조건 없이도 음성 생성 모델과 유사한 동작으로 학습되고 일정 수준의 음성 향상을 수행하는 것을 확인하였고, 이로부터 해당 신경망이 음성 향상에 적합한 구조임을 확인할 수 있다. 신경망에 제약조건을 추가하면 신경망은 더 정확한 음성 생성 모델 동작을 수행하도록 학습되며, 제약조건이 없을 때에 비하여 더 적은 신경망 매개변수로 동등 또는 더 우수한 성능을 제공하는 것을 확인하였다. 또한, 가장 작은 신경망을 사용하는 제안 방법이 Wiener 필터, SEGAN, WaveNet보다 우수한 성능을 가지고, SEGAN과 WaveNet에 비해 각각 1,344배와 70배 적은 신경망 매개변수를 사용하는 것을 확인하였다[7,8,14]. 이를 통해 제안 방법이 음성 생성 모델에 따른 동작을 수행하여 적은 양의 신경망 매개변수로도 효율적인 음성 향상을 수행할 수 있음을 확인하였다.

Ⅱ. 제안하는 음성 향상 방법

1. 개발 배경

신경망을 이용한 대표적인 음성 향상 방법은 시간-주파수 영역에서의 잡음 마스킹(masking)이다[4]. 신경망을 이용하여 입력 신호로부터 마스크를 구하고 입력에 직접 적용하여 원하는 출력 신호를 구한다. 이 방법은 마스크를 기반으로 단순히 입력 신호를 변형하여 품질을 향상시키는 과정에 불과하고, 마스크를 구하는 신경망은 종단간(end-to-end) 학습으로 구한다. 따라서 음성 향상 기술이 신경망 구조와 학습에만 의존하고 음성 특성을 고려한 별도의 동작을 포함하지 않으며, 음성 신호의 고유한 특성을 활용하지 못하는 한계를 가진다. 또한, 음성 특성을 활용한 효율적 동작을 수행하지 못하므로 마스크를 구하는 신경망의 복잡도가 증가하는 경향을 가진다.

본 논문에서는 이와 같은 기존 방법의 한계를 극복하기 위해 음성 생성 모델을 활용하여 매우 효율적으로 음성 향상을 수행하는 방법을 개발하고자 한다. 음성 생성 모델에 따르면 음성 신호는 여기신호와 스펙트럼 포락선의 곱으로 생성되며, 이 모델에 따른 동작을 위해 입력을 여기신호와 스펙트럼 포락선으로 분할하고 각 성분의 품질 향상을 독립적으로 수행한 후에 곱하여 출력을 생성하는 신경망 구조를 설계한다. 이 동작에서 각 성분은 매우 제한된 고유 성질을 가지므로 이를 적극적으로 활용하면 기존 마스킹에 의한 직접적인 신호 변형 방식에 비해 간단한 신경망으로 효율적인 음성 향상이 가능할 것이다.

제안 방법에서는 입력 신호를 여기신호와 스펙트럼 포락선 성분으로 정확히 분할하는 것이 필요하다. 만일 신경망을 종단간 학습하면 분할된 두 성분이 각각 여기신호와 스펙트럼 포락선이 된다는 보장이 없으며, 이 경우 각 성분의 고유 성질을 활용하지 못하므로 원하는 효율적 동작을 기대하기 어렵다. 이를 해결하기 위해 원하는 두 성분으로의 분할을 보장하는 신경망 구조와 학습 방법에 대한 연구가 필요하며, 이 과정을 거쳐 최종 음성 생성 모델에 따라 효율적으로 동작하는 음성 향상 방법을 완성한다.

그동안 머신러닝에서 신경망은 암흑 상자(black box)로 여겨져 왔으며, 신경망 의미를 해석하지 못하면 최적화에 한계를 가진다. 이를 해결하기 위해 해석 가능한 머신러닝 기술(interpretable machine learning)이 최근 높은 관심을 받고 있다[15,16]. 본 논문에서는 음성 생성 모델을 적용하여 신경망 동작의 해석이 가능하도록 하며, 신경망 동작을 구체적으로 분석하고 그 결과를 성능 향상에 활용하도록 한다.

2. 음성 생성 모델 기반의 신경망

음성 생성 모델에 따른 동작을 위한 최소 조건은 두 신호의 곱으로 출력을 생성하는 것이다. 이에 따라 그림 1과 같이 두 개의 모듈이 병렬로 연결된 신경망 구조를 제안한다. 잡음이 포함된 음성신호의 스펙트로그램을 입력하고, 출력단의 활성함수를 제외하고는 완벽하게 동일한 구조를 가지는 두 개의 신경망을 병렬로 배치하고 두 신경망의 출력을 단순히 곱하여 최종 출력 신호를 생성한다. 출력 단의 활성함수에 따라 여기신호 생성 신경망(excitation generator)과 스펙트럼 포락선 생성 신경망(spectral envelope generator)이 결정되며, 다양한 실험을 통하여 활성함수 차이만으로 각 신경망의 동작이 서로 다르고 각각 의도한 여기신호와 스펙트럼 포락선을 생성할 수 있음을 검증하였다. 각 신경

망은 8단의 비인과(non-causal) 1차원 합성곱 계층(1D convolution layer, Conv1D)으로 구성되고, k는 커널(kernel) 크기, c는 채널(channel) 수, a는 활성함수이다. 입력과 출력의 채널 수는 스펙트럼 빈(bin) 개수인 256이며, 그 외 모든 중간 계층의 채널 수는 상수 C로 조절 가능하고 이를 통해 신경망의 크기를 변경한다.

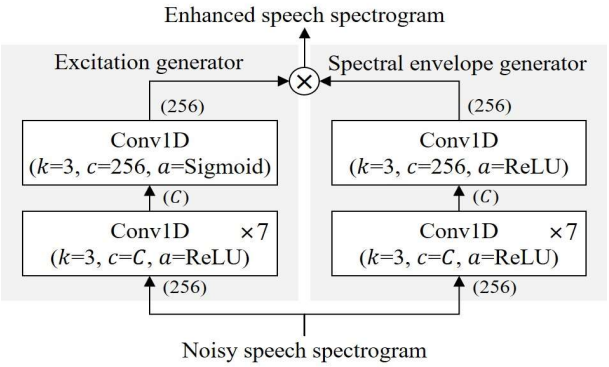


그림 1. 제안하는 신경망 구조
Fig. 1. Architecture of proposed neural network

초기실험에서 그림 1의 신경망을 아무런 외부 조건 없이 종단간 학습시키고, 여기신호와 스펙트럼 포락선으로 분할되어 동작하도록 학습되는지 분석하였다. 그 결과, 단순히 두 신호의 곱으로 출력을 생성하는 구조만으로도 각 신경망이 여기신호와 스펙트럼 포락선을 생성하고 음성 생성 모델에 따라 동작하도록 학습되는 경향이 있음을 확인하였다. 그 이유는 두 신호의 곱으로 음성신호를 생성할 때 여기신호와 스펙트럼 포락선의 곱이 최적 방법이고 신경망이 이를 학습하였기 때문으로 추정된다. 이 결과는 또한 음성 생성 모델이 음성 향상에 적합한 구조임을 간접적으로 보여준다.

그러나 그림 1의 신경망이 항상 여기신호와 스펙트럼 포락선을 추출하지는 못하고 신경망의 매개변수 개수, 초기값, 학습 조건 등에 따라 음성 생성 모델과 다르게 동작하는 경우가 종종 나타나는 것도 확인할 수 있었다. 이에 대한 예는 3장에서 제공한다. 따라서 추가적인 제약조건을 통해 반드시 음성 생성 모델에 따라 동작하도록 학습을 유도하여야 하며, 이를 통해 음성 향상 성능을 확보할 수 있고 해석 가능한 머신러닝의 효과를 기대할 수 있다.

3. 제약조건

그림 1의 신경망이 항상 음성 생성 모델에 따라 동작하도록 학습시키는 방법의 개발이 필요하다. 이를 위한 대표적 방법은 신경망 내부 특정 지점에 독립적인 손실함수(loss function)를 적용하여 원하는 여기신호와 스펙트럼 포락선이 나오도록 하는 것이다. 그러나 이 방법을 사용할 경우 신경망 학습에 3가지 손실함수가 사용되어 학습 효율성이 저하될 수 있고, 손실함수에 필요한 적절한 목표신호를 선택하기 어렵고, 여기신호와 스펙트럼 포락선을 각각 처리하는 과정에서 각 성분의 특성에 따른 효율적 동작을 수행할 수 없다. 따라서 본 논문에서는 복수의 개별적 손실함수 대신에 신경망 자체에 각 성분의 특성에 따른 제약조건을 적용하여 원하는 동작과 효율성을 동시에 얻도록 한다.

그림 1에서 입력 신호는 여기신호와 스펙트럼 포락선 생성을 위한 신경망에 직접 입력된다. 그러나 두 성분은 서로 다른 고유의 성질을 가지므로 제안 방법에서는 각 성질에 적합한 정보만을 입력하는 제약조건을 사용한다. 여기신호를 생성하는 신경망에는 음성신호의 기본 주파수가 존재하는 0 - 1000 Hz 대역의 스펙트로그램만을 입력하고 기본 주파수가 관측 될 경우 이를 바탕으로 유성음 생성에 필요한 배음신호를 생성한다. 또한, 넓은 주파수 대역에 퍼져있는 스펙트럼 포락선 정보를 관측할 수 없게 하여 여기신호를 스펙트럼 포락선으로부터 분리되도록 한다. 스펙트럼 포락선을 생성하는 신경망에는 주파수 영역에서 다운샘플링(down-sampling) 된 스펙트로그램을 입력하여 상대적으로 높은 주파수 해상도를 가지는 여기신호를 관측할 수 없게 한다. 이러한 입력에서의 제약조건을 통해 각각의 신경망은 제한된 정보만을 이용하여 생성할 수 있는 최적의 신호인 여기신호와 스펙트럼 포락선을 생성하게 된다.

그림 2는 제약조건이 적용된 신경망 구조를 보여준다. 여기신호 생성 신경망의 입력정보를 제한하기 위하여 입력 스펙트로그램의 256개 빈 중 저대역에 해당하는 32개 빈만을 잘라(slice) 입력한다. 또한, 스펙트럼 포락선 생성망의 입력정보를 제한하기 위해 입력을 주파수 영역에서 8:1로 다운샘플링 한다. 다운샘플링은 스트라이드(stride)가 적용된 합성곱 연산을 사용하고 커널 크기는 16, 스트라이드는 8이며, 256개 빈을 32개 빈으로 다운샘플링 하고, 커널은

학습을 통해 구한다. 제약조건은 신경망의 입력만 조절하므로 나머지 신경망 구조는 그림 1과 동일하다. 단, 입력되는 스펙트로그램의 차원이 감소하기 때문에 첫 단의 채널 감소로 인해 전체 신경망 매개변수 개수와 연산량이 감소하는 효과를 얻는다.

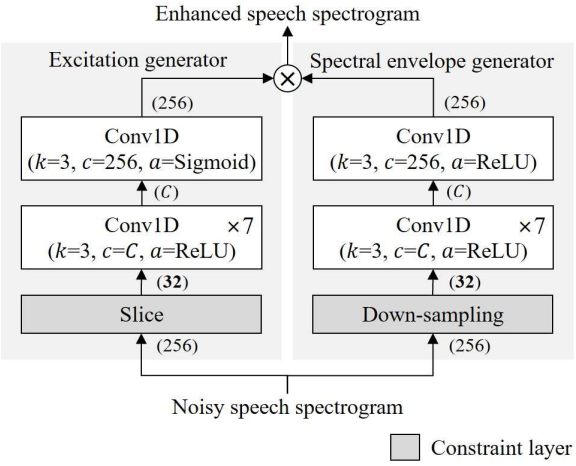


그림 2. 제약조건이 적용된 신경망 구조
Fig. 2. Architecture of neural network with constraints

Ⅲ. 성능평가

1. 데이터셋과 학습 방법

신경망 학습과 성능평가에는 NSDTSEA 데이터셋을 사용하였다[12]. NSDTSEA는 음성 데이터셋 Voice Bank corpus와 다양한 환경 잡음 데이터셋인 DEMAND를 여러 SNR (signal-to-noise ratio)로 결합한 데이터셋이다. Voice Bank corpus에서 30명의 화자를 선정하여 28명은 학습에, 2명은 성능평가에 사용한다. DEMAND에서는 총 13종류의 잡음을 선정하여 8종류는 학습에, 5종류는 성능평가에 사용한다. 추가로 2종류의 인공 잡음신호를 포함하여 총 10종류의 잡음을 학습에 사용한다. 음성신호와 잡음신호를 0, 5, 10, 15 dB SNR로 혼합하여 학습에 사용하고, 2.5, 7.5, 12.5, 17.5 dB SNR로 혼합하여 성능평가에 사용한다. 학습에는 총 11,572개의 신호를 사용하며 성능평가에는 824개의 신호를 사용한다.

NSDTSEA 데이터셋을 16 kHz로 다운샘플링 하고 32,768 샘플(2.048초) 단위로 분리하여 학습에 사용한다. STFT (short-time Fourier transform) 후 절댓값을 취해 스펙트로그램을 구하며, STFT는 512 샘플 사인 윈도우(sine window), 512-포인트 DFT (discrete Fourier transform), 50% 중첩(overlap)을 사용한다. 스펙트럼을 시간 영역 파형으로 변환할 때는 입력 신호의 위상정보를 복제하여 사용한다. 학습 데이터는 9:1로 분리하여 검증(validation) 데이터를 구성하고 조기 종료(early stopping)를 적용하였다.

신경망 구현과 학습에는 Pytorch와 Python을 사용하였다. 다운샘플링을 위한 커널은 0.0625로 초기화 하고 나머지 모든 커널은 He 방법으로 초기화 하였고[17], 미니 배치(mini batch) 크기 16, 평균 절대 오차(mean absolute error) 손실함수, Adam 최적화기를 사용하였다. 학습률 감쇠(learning rate decay)를 적용하여 10 에포크(epoch)마다 0.99배 학습률을 감쇠시켰으며 100 에포크 동안 최저 손실값이 갱신되지 않을 경우 학습을 종료하였다.

성능평가를 위해 총 8개의 신경망을 학습하였고, 각각 신경망 매개변수 개수와 제약조건 여부를 달리 설정하였다. 채널 수 C = 32, 64, 128, 256을 사용하고 C 값에 따라 신경망 매개변수 개수가 결정된다. 그림 1의 각 신경망 이름은 Prop32, Prop64, Prop128, Prop256이며, 제약조건이 적용된 그림 2 신경망 이름 뒤에는 *를 붙였다.

2. 신경망 동작 검증

학습된 신경망이 음성 생성 모델에 따라 동작하는지 검증하기 위해 여기신호와 스펙트럼 포락선 생성 신경망의 각 출력과 두 신호 곱의 스펙트로그램을 분석하였다. 그림 3(a)는 제약조건이 없는 그림 1 신경망의 출력이고, 그림 3(b)는 제약조건이 적용된 그림 2 신경망의 출력이다. 제약조건이 없는 경우에도 신경망이 음성 생성 모델에 따라 동작하려는 경향이 있음을 알 수 있다. 하지만 Prop128과 같이 여기신호와 스펙트럼 포락선 대신에 분석할 수 없는 신호로 학습 되는 경우가 있으며, Prop32와 같이 모호한 여기신호와 스펙트럼 포락선이 생성되는 경우도 있다. 반면, 제약조건을 적용한 경우 신경망 매개변수 개수에 상관없이 항상 음성 생성 모델에 따라 동작하는 것을 알 수 있다. 이

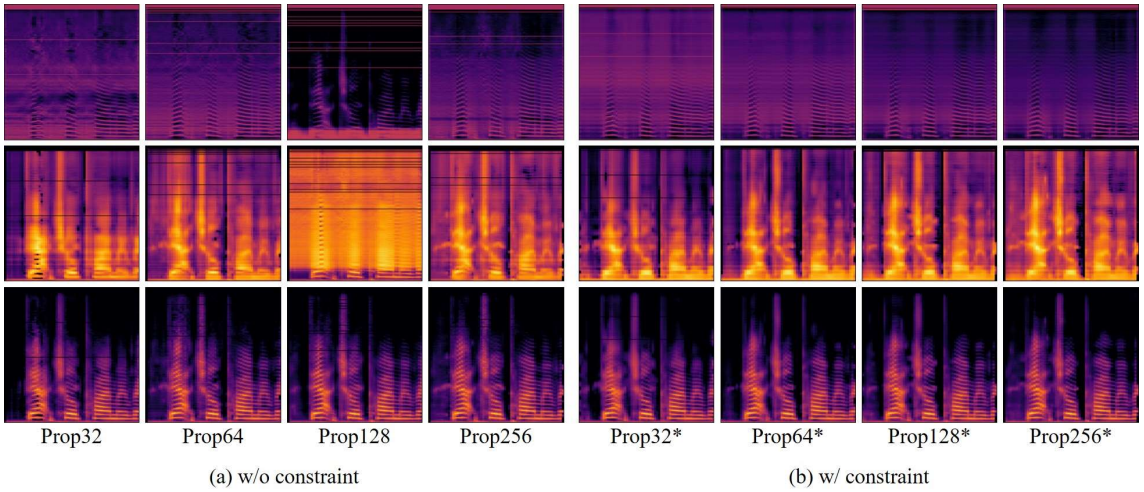


그림 3. 제약조건 여부에 따른 스펙트로그램 비교, (위) 여기신호, (가운데) 스펙트럼 포락선, (아래) 두 신호의 곱
Fig. 3. Neural network output depending on the constraint condition, (top) excitation signal, (middle) spectral envelope, (bottom) product of the two signals

를 통해 제안하는 제약조건이 음성 생성 모델에 따라 동작하도록 신경망 학습을 유도하는 역할을 하고, 원하는 음성 향상 동작을 위해 필요한 조건임을 알 수 있다.

그림 4는 Prop128*의 스펙트로그램 비교이다. 잡음이 포함된 음성신호로부터 깨끗한 음성의 여기신호와 스펙트럼 포락선을 각각 생성하고 두 신호의 곱인 최종 출력에서 잡음이 제거된 음성을 얻는다. 여기신호를 보면 유성음 구간에서 배음신호가, 무성음 구간에서 백색 잡음신호가 생성되었음을 확인할 수 있다. 이를 통해 제안 방법이 음성 생성 모델에 따라 동작하여 여기신호와 스펙트럼 포락선의 품질 향상을 각각 수행하여 최종 출력을 생성하는 것을 확인할 수 있다.

제안 방법의 스펙트로그램을 Wiener 필터, SEGAN, WaveNet의 공개된 음원 스펙트로그램과 비교하였다[18,19]. 그림 5는 각 방법에 대한 스펙트로그램이고, (a) 영역을 보면 Prop128*이 제약조건에 의해 기본 주파수 대역으로부터 배음신호를 생성했기 때문에 다른 방법보다 깨끗한 신호를

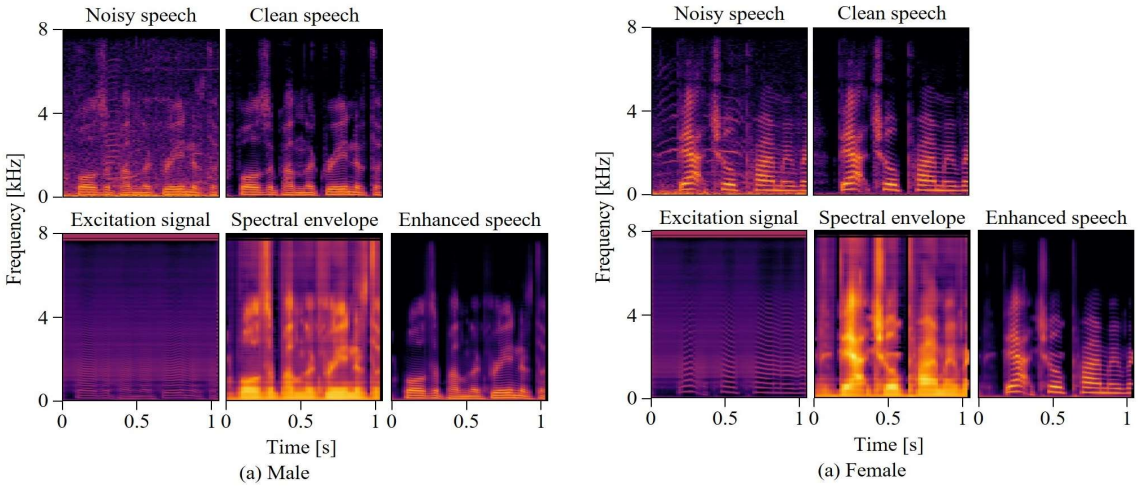


그림 4. Prop128*의 스펙트로그램
Fig. 4. Spectrogram for Prop128*

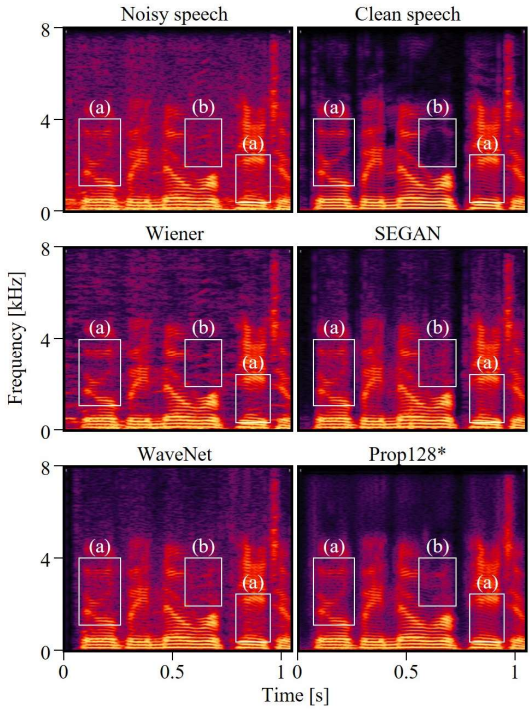


그림 5. 각 방법별 스펙트로그램[18,19]
Fig. 5. Spectrogram for each method

출력한 것을 확인할 수 있다. (b) 영역에서 다른 방법들은 톤(tone) 성분의 잡음신호와 유성음의 배음신호를 구분하지 못하여 낮은 잡음제거 성능을 가지지만, 제안방법은 기본 주파수 대역으로부터 배음신호를 생성했기 때문에 배음신호가 (b) 영역에 존재하는 잡음신호를 포함하지 않으므로 출력에서 해당 잡음신호를 제거할 수 있음을 확인할 수 있다. 즉, 제약조건으로 인해 배음신호 생성은 1000 Hz 이하의 정보만을 사용하므로 고대역에 존재하는 잡음 정보를 활용하지 않고, 따라서 해당 잡음신호를 배음신호 생성에 필요한 하모닉 성분으로 오인하지 않고 깨끗한 음성의 배음신호를 생성할 수 있다.

3. 객관적 성능 평가

객관적 성능 평가는 PESQ, CSIG, CBAK, COVL을 통해 진행하였다[13]. PESQ는 인간의 청각인지 능력을 고려한 음성 신호의 품질 평가 지표이다. CSIG는 신호 왜곡 정도에 대한 MOS (mean opinion score) 추정치, CBAK은 배경잡음 왜곡 정도에 대한 MOS 추정치, COVL은 전반적인 품질에 대한 MOS 추정치이다.

표 1은 신경망 매개변수 개수, 제약조건 여부에 따른 객관적 성능평가 결과이며, 입력 SNR 2.5, 7.5, 12.5, 17.5 dB에 대한 각 성능의 평균이다. C가 128 이하에서 신경망 매개변수 개수에 비례하여 성능이 증가하지만, 그 이후는 성능 향상이 없다. 제약조건 여부에 따른 성능 차이를 보면, 모든 경우에서 제약조건이 적용된 경우 더 적은 신경망 매개변수로 동등 또는 더 우수한 성능을 보이는 것을 알 수 있다. 이는 제약조건을 통해 신경망에 필요한 정보만을 제공함으로써 학습 과정에서 국소지점에 빠지지 않고 효율적으로 매개변수를 사용했기 때문이라고 판단된다. 단, 제약조건이 적용되지 않은 경우 학습 조건, 초기값 등의 변경에 따라 음성 생성 모델에 따른 동작을 보장할 수는 없다. 본 논문의 결과는 여러 조건으로 학습한 결과 중 하나이며, 재현 시 다른 경향으로 학습될 수 있다. 따라서 제약조건이 없는 모델에서 음성 생성 모델에 따른 동작이 학습 안 될 경우 성능 평가 결과 또한 변경될 수 있다.

표 2는 입력 SNR에 따른 Prop128*의 성능이다. 모든 입력 SNR에서 정상 동작을 수행하며, 입력 SNR에 비례하여 성능이 증가하는 것을 보여준다.

표 1. 제안한 방법의 객관적 성능평가 결과
Table 1. Objective evaluation results of proposed method

Model	PESQ	CSIG	CBAK	COVL	# of parameters
Noisy input	1.97	3.35	2.44	2.63	-
Prop32	2.49	3.69	3.07	3.08	0.14M
Prop32*	2.49	3.75	3.07	3.11	0.09M
Prop64	2.65	3.93	3.19	3.28	0.35M
Prop64*	2.63	3.96	3.18	3.29	0.26M
Prop128	2.66	3.91	3.20	3.28	0.99M
Prop128*	2.69	4.03	3.20	3.36	0.81M
Prop256	2.65	3.92	3.19	3.29	3.15M
Prop256*	2.69	4.00	3.18	3.34	2.81M

표 2. 입력 SNR에 따른 Prop128*의 성능평가 결과
Table 2. Objective evaluation results of Prop128* depending on input SNR

Input SNR	PESQ	CSIG	CBAK	COVL
2.5	2.16	3.49	2.77	2.81
7.5	2.61	3.97	3.13	3.29
12.5	2.84	4.20	3.33	3.52
17.5	3.14	4.46	3.59	3.81

제안한 음성 향상 방법의 성능과 복잡도를 동일한 데이터셋에 대해 성능이 공개되어 있는 Wiener 필터, SEGAN, WaveNet과 비교하였고[9], 표 3에 결과가 정리되어 있다. WaveNet CBAK 항목을 제외하면, 가장 적은 신경망 매개변수를 사용하는 Prop32*가 비교 방법보다 우수한 성능을 가진다. 다만, CBAK은 추정한 잡음 신호의 왜곡을 측정하는 지표로 음성 향상 성능에서 핵심 지표는 아니다. 신경망 매개변수의 개수를 비교하면, Prop32*가 SEGAN와 WaveNet에 비해 각각 1,344배와 70배 적은 매개변수를 가지며, 가장 우수한 성능을 가지는 Prop128*는 각각 149배와 8배 적은 신경망 매개변수를 가진다.

표 3. 각 방법의 객관적 성능 비교
Table 3. Objective performance comparison of each method

Model	PESQ	CSIG	CBAK	COVL	# of parameters
Noisy input	1.97	3.35	2.44	2.63	-
Wiener	2.22	3.23	2.68	2.67	-
SEGAN	2.16	3.48	2.94	2.80	121M
WaveNet	-	3.62	3.23	2.98	6.31M
Prop32*	2.49	3.75	3.07	3.11	0.09M
Prop64*	2.63	3.96	3.18	3.29	0.26M
Prop128*	2.69	4.03	3.20	3.36	0.81M

Ⅳ. 결론

본 논문에서는 음성 생성 모델을 이용하여 음성신호에 특화된 신경망 구조를 설계하여 음성 향상을 효율적으로 수행하는 새로운 머신러닝 기반의 음성 향상 기술을 제안하였다. 음성 생성 모델에 따라 여기신호와 스펙트럼 포락선의 곱으로 출력을 구하는 신경망을 설계하였고, 여기신호와 스펙트럼 포락선의 고유 성질에 따른 제약조건을 신경망에 적용하여 음성 생성 모델에 따른 동작 학습을 강화하고 동작의 복잡도를 감소시켰다. 제안 방법 중에서 가장 작은 신경망이 Wiener 필터, SEGAN, WaveNet 보다 우수한 객관적 성능을 가지고, 신경망 매개변수가 SEGAN과 WaveNet에 비해 각각 1,344배와 70배 적다. 이를 통해 제안한 신경망과 제약조건을 사용하여 음성 생성 모델에 따라 동작하는 새로운 음성 향상 방법이 가능하고 기존 방법에 비해 매우 작은 신경망으로 효율적인 음성 향상이 가능한 것을 검증하였다.

참 고 문 헌 (References)

[1] J. Jung and G. Kim “Adaptation of classification model for improving speech intelligibility in noise,” *J. of Broadcast Engineering*, Vol. 23, No. 4, pp. 511-518, Jul. 2018.
doi: https://doi.org/10.5909/JBE.2018.23.4.511

[2] G. Kim “A post-processing for binary mask estimation toward improving speech intelligibility in noise,” *J. of Broadcast Engineering*, Vol. 18, No. 2, pp. 311-318, Mar. 2013.
doi: https://doi.org/ 10.5909/JBE.2013.18.2.311

[3] Y. Wang, “Research progress in speech enhancement technology,” *Int. Conf. on CVIDL*, Chongqing, China, pp. 222-226, 2020.
doi: https://doi.org/ 10.1109/CVIDL51233.2020.00-97

[4] X. Du, M. Zhu, X. Shi, X. Zhang, W. Zhang, and J. Chen, “End-to-end model for speech enhancement by consistent spectrogram masking,” *arXiv preprint, arXiv:1901.00295*, 2019.
doi: https://doi.org/ 10.48550/arXiv.1901.00295

[5] 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, Nov. 2022.
doi: 10.5909/JBE.2022.27.6.840

[6] C. Macartney and T. Weyde, “Improved speech enhancement with the Wave-U-net,” *arXiv preprint arXiv:1811.11307*, 2018.
doi: https://doi.org/ 10.48550/arXiv.1811.11307

[7] D. Rethage, J. Pons, and X. Serra, “A Wavenet for speech denoising,” *Proc. of ICASSP*, Calgary, AB, Canada, pp. 5069-5073, 2018.
doi: https://doi.org/ 10.1109/ICASSP.2018.8462417.

[8] S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: Speech enhancement generative adversarial network,” *Proc. of Interspeech*, Stockholm, Sweden, pp. 3642-3646, 2017.
doi: https://doi.org/10.21437/Interspeech.2017-1428

[9] J. Su, Z. Jin, and A. Finkelstein, “HiFi-GAN: high-fidelity denoising and dereverberation based on speech deep features in adversarial networks,” *arXiv preprint arXiv:2006.05694*, 2020.
doi: https://doi.org/ 10.48550/arXiv.2006.05694.

[10] M. Soni, N. Shah and H. Patil, “Time-frequency masking-based speech enhancement using generative adversarial network,” *Proc. of ICASSP, Calgary*, AB, Canada, pp. 5039-5043, 2018.
doi: https://doi.org/ 10.1109/ICASSP.2018.8462068.

[11] L. Raphael, G. Borden, and K. Harris, *Speech science primer*, (K. Kim, Trans.), Williams & Wilkins, Philadelphia, pp. 83-148, 2004.

[12] 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/10.7488/ds/2117.

[13] Evaluation measures open source, https://www.crcpress.com/down

loads/K14513/K14513_CD_Files.zip (accessed Nov. 1, 2022).
[14] P. Scalart and J. V. Filho, “Speech enhancement based on a priori signal to noise estimation,” *Proc. of ICASSP*, Atlanta, GA, USA, pp. 629-632 vol. 2, 1996.
doi: https://doi.org/10.1109/ICASSP.1996.543199.
[15] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)”, *IEEE Access*, vol. 6, pp. 52138-52160, 2018.
doi: https://doi.org/10.1109/ACCESS.2018.2870052.
[16] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, “Interpretable machine learning: fundamental principles and 10 grand challenges,” *arXiv preprint, arXiv:2103.11251*, 2021.
doi: https://doi.org/ 10.48550/arXiv.2103.11251
[17] K. He, X. Zhang, S. Ren, J. Sun “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” *Proc. of ICCV*, Santiago, Chile, 2015, pp. 1026-1034.
doi: https://doi.org/ 10.1109/ICCV.2015.123.
[18] Wavenet demo source, http://www.jordipons.me/apps/speech-denoising-wavenet (accessed Sep. 1, 2023).
[19] SEGAN demo source, http://veu.talp.cat/segan (accessed Sep. 1, 2023).

저 자 소 개

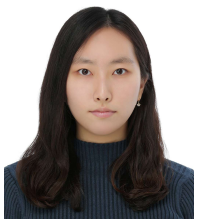
유 정 찬

- 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

김 재 원

- 2019년 2월 : 광운대학교 전자공학과 학사
- 2019년 3월 ~ 현재 : 광운대학교 전자공학과 석박통합과정
- ORCID : https://orcid.org/0000-0002-6496-842X
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

문 희 연



- 2023년 2월: 광운대학교 전자공학과 공학사
- 2023년 3월 ~ 현재: 광운대학교 전자공학과 석사과정
- ORCID : https://orcid.org/0009-0003-3363-9775
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

저 자 소 개

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