간이 보청기를 활용한 TasNet-GAN 모델 기반 오디오 노이즈 캔슬링에 관한 연구

서성윤, 말체르빠 아나스타시아, 김우휘, 안젤라 에밀 조세, 권오송, 김재수*

경북대학교

kesa0v0@gmail.com, 1803nastiama2003@gmail.com, rladngnl@gmail.com, angelaemilejose@gmail.com, osong030110@gmail.com, kjs@knu.ac.kr*

A Study on Audio Noise Cancellation Using a TasNet-GAN Model with a Simple Hearing Aid

SeongYun Seo, Malysheva Anastasia, WooHwi Kim, Angela Emile Jose, OSong Kwon, JaeSoo Kim

Kyungpook Univ

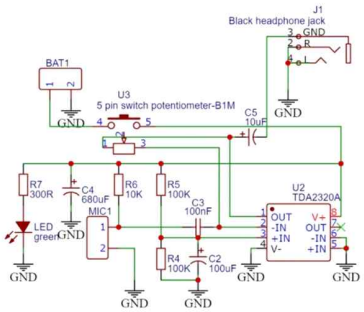
약

논문은 저비용·저전력으로 구현 가능한 간단한 보청기 회로를 설계하고, 딥러닝 기반 신경망 모델을 적용하여 출력 음성의 품질을 향상시키는 방법을 제안한다. 회로는 기본적인 음성 증폭 기능을 수행하며, 모델은 생성적 적대 신경망(GAN)과 Conv-TasNet 구조를 사용하여 잡음과 왜곡을 제거하였다. 학습 과정에서 Log-Mel 손실 함수를 결합하여 복원 성능을 최적화하였고, 성능은 PESQ와 STOI 지표를 통해 평가하였다. 그 결과, PESQ는 약 31%, STOI는 14% 향상되었으며, 모델 경량화를 통해 제한된 하드웨어 환경에서도 안정적인 음성 향상이 가능함을 보여준다.

Ⅰ.

전 세계적으로 많은 인구가 청각 장애를 겪고 있으며, 청력 손실은 시각·운동·정신 장애에 이어 네 번째로 흔한 장애로 알려져 있다. 인공 와우(Cochlear Implant)는 청력 손실을 개선하는 효과적인 기술로 평가되지만, 높은 비용과 수술 위험성으로 인해 일반 사용자의 접근성이 제한적이다. 이에 따라 구조가 단순하고 저비용·저전력으로 동작하는 보청기 개발의 필요성이 제기되고 있다. 한편 최근 음향 신호처리 분야에서는 딥러닝 기반 신경망 모델이 잡음 제거와 음질 향상에서 우수한 성능을 보이는 것으로 보고되고 있으며[1], 특히 생성적 적대 신경망(GAN)과 Conv-TasNet 구조는 시간 도메인 파형을 직접 처리하여 음성의 왜곡을 최소화할 수 있는 장점을 가진다. 따라서 본 연구에서는 저비용·저전력으로 구현 가능한 간단한 보청기 회로를 설계하고, 여기에 딥러닝 기반 신경망 모델을 결합하여 출력 음성의 품질을 향상시키는 방법을 제안한다. 본 연구는 하드웨어 회로와 인공지능 알고리즘의 융합을 통해 기존 보청기의 비용과 성능 한계를 개선하고, 향후 스마트 보청기 개발의 가능성을 제시하는 데 목적이 있다.

Ⅱ. 해결 방법



(그림 1) 간이 보청기 회로도

Ⅱ-1. 간이 보청기 설계

(그림 1)은 실험에 사용한 간이 보청기 회로도 이다. 간이 보청기 하드웨어는 외부 음성을 전기 신호로 변환하는 마이크 입력부, 신호를 증폭하는 오디오 증폭부, 음량을 조절하고 소리를 출력하는 제어 및 출력부로 구성된다. 핵심 부품으로는 저전력 듀얼 오디오 증폭기인 TDA2320A(U2)를 사용하였다. 이 IC는 적은 외부 부품으로도 안정적인 증폭이 가능하며 소형 회로 설계에 적합하다. 마이크(MIC1)를 통해 입력된 음성 신호는 증폭기(U2)에서 증폭되고, 5핀 스위치 전위차계(U3)를 통해 사용자가 전원 ON/OFF 및 음량을 조절할 수 있다. 최종 증폭된 신호는 헤드폰 잭(J1)을 통해 출력된다.

Ⅱ-2. 신경망을 이용한 음향 품질 최적화

본 연구에서는 하드웨어 시스템의 성능적 한계를 보완하기 보다 효율적인 음성 복원을 실현하기 위해 딥러닝 기반 신경망 모델을 도입하였다. 본 모델의 주요 목적은 음성 증폭 과정에서 발생하는 잡음과 왜곡을 제거하여 원음에 가까운 청명한 음성 신호를 복원하는 것이다. 이를 위해 연구자는 AIHub에서 제공하는 공개 음성 데이터를 활용하였으며, 고성능 마이크로 직접 녹음한 기준(청정) 음성과 간이 보청기를 이용해 녹음한 잡음이 포함된 음성을 결합하여 학습용 데이터셋을 구축하였다.

제안된 모델은 생성적 적대 신경망(Generative Adversarial Network, GAN) 구조를 기반으로 하며, 생성기(Generator) 와 판별기(Discriminator) 로 구성되어 상호 경쟁적인 학습을 수행한다. 생성기는 시간 도메인(Time-domain)에서 음성 분리와 향상에 탁월한 성능을 보이는 Conv-TasNet 아키텍처를 채택하였으며, 파형 신호를 직접 입력과 출력으로 사용함으로써 위상 정보를 유지하고 음성의 시간적 및 주파수적 특성을 효율적으로 모델링할 수 있다.

모델은 1D Convolution 기반의 인코더(Encoder), 다수의 시간 컨볼루션 네트워크 블록(TCN Block) 으로 구성된 분리·향상 모듈(Separation Module), 그

리고 향상된 특성을 다시 파형 형태로 복원하는 디코더(Decoder) 로 이루어진다. 한편, 판별기(Discriminator) 는 생성된 음성과 실제 음성을 구분하는 역할을 하며, 학습의 안정성을 보장하기 위해 스펙트럼 정규화(Spectral Normalization) 가 적용된 1D Convolution 블록으로 설계되었다. 판별기는 입력된 파형에 대해 진위 판정 점수를 산출함으로써 생성기의 학습 품질을 향상시킨다.

모델 학습 과정에서는 세 가지 손실 함수를 함께 사용하였다. 첫째, SI-SDR(Scale-Invariant Signal-to-Distortion Ratio) 손실 함수는 파형 간의 유사도를 정량적으로 측정한다. 둘째, MR-STFT(Multi-Resolution Short-Time Fourier Transform) 손실 함수는 다양한 해상도에서 스펙트럼 구조를 유지하도록 돕는다. 셋째, Log-Mel 지각 손실(Log-Mel Perceptual Loss) 은 청각적 품질을 개선하고 음성의 지각적 명료도를 향상시키기 위해 사용된다.

이와 같은 학습 구조를 통해 제안된 모델은 제한된 하드웨어 환경에서도 효율적이고 안정적인 음성 향상이 가능하도록 설계되었다.

Ⅲ. 비교 및 분석

Ⅲ-1 제안 모델의 성능 평가

훈련된 모델의 객관적 성능을 검증하기 위해 학습에 사용되지 않은 독립적인 테스트 데이터셋을 활용하여 평가를 수행하였다. 성능 평가는 음성의 주관적 품질과 명료도를 측정하기 위한 두 가지 객관적 지표를 중심으로 이루어졌다. 첫 번째 지표인 PESQ(Perceptual Evaluation of Speech Quality)는 청취자가 인지하는 음질 수준을 평가하기 위한 기준이며, 두 번째 지표인 STOI(Short-Time Objective Intelligibility) 는 음성의 명료도와 전달력을 정량적으로 측정하는 지표이다. 평가 절차는 노이즈가 포함된 원본 음성(Noisy)과 모델이 출력한 향상된 음성(Denoised)을 각각 깨끗한 기준 음성과 비교하는 방식으로 진행되었다. <표 1>은 제안된 모델의 성능 향상 결과를 요약한 것으로, 평균적으로 PESQ 점수는 1.4936에서 1.9658로 약 31% 향상, STOI 점수는 0.6958에서 0.7945로 약 14% 향상되는 결과를 보였다.

<표 2>는 GAN 기반의 대표적인 음질 향상 모델인 MetricGAN(Fu et al., 2019)[2] 논문 결과와 성능을 비교하였다. 제안한 모델과 비교대상으로 선정된 MetricGAN 모델 논문은 동일한 음성 향상 문제를 다루고 있으나, 서로 다른 데이터셋(TIMIT)을 사용하였다. 따라서 직접적인 수치 비교에는 한계가 있으나, PESQ 및 STOI 성능 지표는 음성 품질과 명료도 개선의 상대적 척도로 참고할 수 있다.<표 2>의 PESQ 점수의 경우, 제안 모델(TasNet-GAN)은 MetricGAN(P)보다 약 0.167 낮게 측정되었지만 이는 MetricGAN이 PESQ와 같은 객관적 평가지표 자체를 최적화 목표로 삼는 학습 전략을 사용하기 때문으로 분석된다. 음성의 명료도를 나타내는 STOI 점수에서는 제안 모델이 0.7945를 기록하며 비교 모델보다 우수한 성능을 보였다.

<표 1> PESQ, STOI 성능 평가 결과

평가지표	처리 전(Noisy)	처리 후(Denoised)
PESQ	1.4936	1.9658
STOI	0.6958	0.7945

< 표 2 > 타 모델과 비교

모델	PESQ	STOI
제안모델(TasNet-GAN)	1.9658	0.7945
MetricGAN(P)	2.133	0.760
MetricGAN(S)	2.025	0.768

Ⅲ-2 경량화 모델의 성능 비교

<표 3>은 제안모델의 경량화 된 버전(TasNet-GAN light)의 성능 평가 결과이다. PESQ는 1.7718, STOI는 0.7662로 두 가지 성능 지표 모두 제안 모델에 비해 성능이 다소 감소하였다. 하지만 기존 제안 모델에 비해 파라미터 수가 생성자 25,880,648개에서 993,678개로 판별자 82,657개에서 81,953로 크게 줄어들었다. PESQ 값은 다소 감소하였지만 STOI 값은 7.662로 제안 모델(0.7945), MetricGAN(S)(0.768)과 거의 대등한 수준을 유지했다. 제안 모델(TasNet-GAN)은 융합형 네트워크 모델로, 경량화되지 않은 버전이며 MetricGAN은 성능은 뛰어나지만 파라미터 수가 너무 많아 보청기에는 적용하기 어렵다. TasNet-GAN Light는 경량화된 버전으로, 성능이 다소 감소하였으나 비교 모델에 비해 제한된 하드웨어 환경에서 효율적으로 활용할 될 수 있어 실용성이 높다.

<표 3> 경량화 모델 성능 평가 결과

평가지표	처리 전(Noisy)	처리 후(Denoised)
PESQ	1.4936	1.7718
STOI	0.6958	0.7662

IV. 결론

본 연구에서는 저비용으로 구현 가능한 간이형 보청기 시스템을 설계하고, 딥러닝 기반 신경망 모델을 활용하여 출력 음성의 음질을 향상시키는 방법을 제안하였다. 실험에서는 TDA2320A 오디오 증폭기를 중심으로 회로를 구성하였으며, 시간 도메인 파형을 직접 처리할 수 있는 Conv-TasNet 구조의 생성적 적대 신경망(GAN) 을 적용하여 음성 신호를 향상시켰다. 실험 결과, 제안된 모델은 처리 전 음성과 비교하여 PESQ 약 31%, STOI 약 14%의 성능 향상을 보였으며, 이는 주관적 음질과 음성 명료도 모두에서 유의미한 개선을 달성했음을 의미한다. 이러한 결과는 소프트웨어적 신경망 처리를 통해 아날로그 회로의 한계를 보완할 수 있음을 입증하였으며, 간단한 구조의 보청기에서도 충분히 만족스러운 청각적 품질을 구현할 수 있음을 보여준다. 향후 연구에서는 제안된 모델에 지식 증류(Knowledge Distillation)와 같은 경량화 기법을 적용하여, 성능을 유지하면서도 파라미터 수와 계산 부담을 줄이는 방향으로 발전시킬 수 있으며 제안된 모델을 경량화 및 임베디드 시스템에 최적화하여 실시간 음질 개선이 가능한 시스템으로 확장함으로써, 다양한 청취 환경에서도 잡음을 효과적으로 억제하고 청각 장애인의 생활 품질을 실질적으로 향상시킬 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

" 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2021-0-01082)

이 연구는 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 "한국어 음성"을 활용하여 수행된 연구입니다. 본 연구에 활용된 데이터는 AI 허브(aihub.or.kr)에서 다운로드 받으실 수 있습니다.

고 문 헌

[1] Le, H.-T., et al., "A Review of Deep Learning Techniques for Speech Processing," arXiv preprint arXiv:2401.03339, 2024.

[2] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR, vol. 97, pp. 2031–2041, 2019.

[3] Luo, Y. and Mesgarani, N., "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 27, no. 8, pp. 1256-1266, Aug. 2019.