Journal of KIIT. Vol. 24, No. 2, pp. 25-33, Feb. 28, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.2.25

전장환경에의 적용 가능성 검증을 위한 소음 제거 모델 및 음성인식 모델의 성능 비교 연구

김동훈*, 안서경**, 이수진***

Performance Comparison of Denoising Models and Speech Recognition Models for Verifying the Applicability to Battlefield

Donghoon Kim*, Seokyeong An**, and Soojin Lee***

요 약

음성인식 기술은 인간과 기계의 원활한 상호작용을 보장하는 핵심 수단으로 부상하고 있다. 그러나 현재의 음성인식 모델들은 충격성 소음이 지배적인 전장환경에서의 적용 가능성이 검증되지 않았다. 이에 본 연구에서는 음성인식 모델의 전장환경 적용에 필수적인 소음 제거 모델 4종을 대상으로 효율성을 확인하고, 3종의 음성인식 모델들과의 결합을 통한 성능 개선 효과를 비교 분석하였다. 전장환경을 모사하는 데이터세트는 일반인의 발화 데이터와 전장 소음 데이터를 3단계의 SNR로 합성하여 구축하였다. 실험 결과, 선정된 소음 제거 모델들은 모두 실시간 처리 능력을 가지며, Fullsubnet 모델과 Wav2Vec 2.0 모델을 결합하여 파인튜닝하는 방식이 가장 극심한 소음 환경에서도 강건한 성능을 달성함을 확인하였다.

Abstract

Speech recognition technology is emerging as a key tool for ensuring seamless human-machine interaction. However, current speech recognition models have not yet been proven to be applicable to battlefield dominated by impulsive noise. Therefore, this study examined the effectiveness of four denoising models, essential for applying speech recognition models to battlefield, and compared and analyzed the performance improvements achieved by combining them with three speech recognition models. An experimental dataset simulating battlefield environments was constructed by synthesizing human speech data and battlefield noise data at three Signal-to-Noise Ratios (SNR). Experimental results demonstrated that all selected denoising models possess real-time processing capabilities, and fine-tuning approach combining FullSubNet model with Wav2Vec 2.0 model achieved the most robust performance even in the extreme noise environments.

Keywords

speech recognition, battlefield noise, denoising model, signal to noise ratio

* 육군3사관학교 전자공학과 강사
- ORCID: https://orcid.org/0009-0005-9945-0558
** 육군본부 지상군페스티벌기획단 사이버대응장교
- ORCID: https://orcid.org/0009-0001-9102-2613
*** 국방대학교 사이버·컴퓨터공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-4117-407X

· Received: Nov. 19, 2025, Revised: Dec. 12, 2025, Accepted: Dec. 15, 2025

· Corresponding Author: Soojin Lee
Dept. of Cyber Security and Computer Engineering, Korea National Defense University, 1040, Hwangsanbeol-ro, Nonsan-si, Chungcheongnam-do, Republic of Korea
Tel.: +82-41-831-5378, Email: cyberkma@korea.kr

Ⅰ. 서 론

미래 전장환경에서는 AI 기반 군사참모나 유무인 복합체계(MUM-T, Manned-Unmanned Teaming) 등과 같이 인간과 기계의 원활한 상호작용이 요구되는 체계가 활발하게 운영될 것이다. 그리고 인간 운영자와 기계 간의 상호작용에서는 음성인식 기술이 핵심적인 수단이 될 것이다. 그러나 체계 운영자나 전투원이 음성을 통해 각종 장비를 정확하게 제어하고 체계 활용성을 극대화하기 위해서는 신뢰할 수 있는 음성인식 기술의 확보가 중요하다.

최근 자동음성인식(ASR, Automatic Speech Recognition) 모델의 성능이 크게 개선됨에 따라 음성인식 기술의 적용도 다방면으로 확장되고 있다. 그러나 국방 분야, 특히 군사작전 수행과 관련된 환경 하에서의 직접적인 적용 사례는 확인하기 어렵다. 이와 같이 ASR 모델의 전장환경 적용이 제한되는 원인으로는 모델이 군사용어를 학습하지 않아 인식률이 현저하게 떨어진다는 점을 들 수 있다. 그리고 가장 심각한 문제는 대부분의 모델이 일반적인 소음 환경에서는 우수한 성능을 보이는 반면, 총성 및 포탄 소리 등 충격성 소음이 지배적인 전장환경에서는 인식률이 심각하게 저하된다는 점이다. 따라서 ASR 모델의 전장 활용을 활성화하기 위해서는 극심한 소음이 발생하는 상황에서도 음성을 명확하게 인식할 수 있는 강건성이 보장되어야 한다.

ASR 모델의 강건성을 강화하기 위한 접근법은 크게 두 가지로 구분할 수 있다. 첫 번째 접근법은 음성을 소음 제거 모델(Denoising model)을 통해 먼저 정제한 후 ASR 모델에 전달하는 방식이다. 두 번째 접근법은 모델 학습 단계에서 소음이 포함된 음성 데이터로 모델 자체를 학습(Fine-tuning)시켜 소음에 직접적으로 적응시키는 방식이다.

이러한 접근법과 관련된 선행 연구들은 다양한 아키텍처를 가진 소음 제거 모델과 ASR 모델을 일반적인 소음 환경에 적용하여 효과성을 입증하였다. 그러나 전장환경과 같은 극단적인 소음 환경에서 소음 제거 모델의 효율성과 ASR 모델의 최종 인식 성능 개선 효과를 분석한 연구는 진행된 바가 없다.

이에 본 연구에서는 전장환경의 특수성을 최대한 반영하여, 광범위하게 활용되는 소음 제거 모델의 효율성과 ASR 모델에 대한 인식 성능 개선 효과를 비교 분석하였다. 분석 대상은 서로 다른 구현 방식을 가진 소음 제거 모델 4종(FullSubNet[1], U-Net[2], Wave-U-Net[3], SEGAN[4])과 ASR 모델 3종(Whisper tiny, Whisper small[5], Wav2Vec 2.0[6])을 선정하여 비교하였다.

본 연구의 기여는 다음과 같다. 첫째, 기존 연구 대비 보다 현실적인 전장 소음 환경을 반영한 데이터세트를 구축하여 소음 제거 모델과 ASR 모델의 성능을 정량적으로 검증하였다. 둘째, 단순히 모델 간의 성능 비교에 그치지 않고, 모델의 구조적 특성 및 효율성까지 분석하여 전장환경에서의 적용 가능성을 검증하였다. 셋째, 실시간 처리와 정확성이 요구되는 전장환경의 특수성을 고려하여, 소음 제거 모델과 ASR 모델의 효율적인 결합방식을 모색하고 실제 작전 운용이 가능한 실증적인 방안을 제시하였다.

본 논문의 구성은 다음과 같다. 2장에서는 분석 대상으로 선정한 소음 제거 모델 및 ASR 모델에 대해 설명하고, 음성인식 성능 개선과 관련된 선행 연구를 정리한다. 3장에서는 전장환경에서의 소음을 모사한 데이터세트의 구축 및 실험 방법을 설명한 후, 4장에서는 실험 환경과 성능 평가 결과를 정리한다. 마지막으로 5장에서는 연구 결과를 요약하고 결론을 제시한다.

Ⅱ. 관련 연구

2.1 소음 제거 모델

2.1.1 FullSubNet

FullSubNet 모델[1]은 마스킹(Masking) 기반 모델의 실시간 음성 향상을 목표로 제안되었으며, 음성의 시간적 맥락을 모델링하기 위해 LSTM(Long Short-Term Memory)을 활용한다. 음성신호를 전체 주파수 대역(Full-band)과 여러 개의 부대역(Sub-band)으로 구분해 동시에 처리한 뒤, 그 결과를 융합하여 정교한 마스크를 예측한다. 이후 예측된 마스크를 소음이 섞인 스펙트럼에 곱하여 소음 부분만을 선택적으로 억제하는 필터링(Filtering)을 수행한다.

2.1.2 U-Net

생성(Generation) 기반 모델의 대표적인 사례라고 할 수 있는 U-Net 모델[2]은 처음에는 의료분야에서 영상을 분할하는 목적으로 제안된 모델이다. 대칭적 인코더-디코더 구조를 가지고 있으며, 인코더의 각 계층에서 추출된 특징 맵을 스킵 연결(Skip connection)을 통해서 디코더의 해당 계층으로 직접 전달한다. 이러한 U-Net 모델의 구조가 음성신호의 시간-주파수 특성을 보존함과 동시에 소음을 분리하는 데 효과적이라는 사실이 알려지면서 음성인식 성능 향상 분야에서도 널리 사용되고 있다.

2.1.3 Wave-U-Net

Wave-U-Net 모델[3]은 U-Net 아키텍처를 1차원 음성 파형 데이터에 직접 적용하도록 변형한 모델이다. U-Net과 동일하게 스킵 연결을 활용하여 다운샘플링 과정에서 손실될 수 있는 원본 신호의 세부적인 시간 정보를 보존함으로써 음성 신호의 왜곡을 최소화한다. 이 모델은 기존 모델들이 스펙트럼 변환 과정에서 필연적으로 겪는 위상(Phase) 정보 손실을 해결하기 위해 파형을 직접 처리하는 것이 핵심적인 특징이다. 스펙트럼 변환 시 원본 파형을 인코더를 통해 직접 다운샘플링하여 특징을 추출하고, 이를 다시 디코더를 통해 업샘플링하여 깨끗한 파형을 복원한다.

2.1.4 SEGAN

SEGAN(Speech Enhancement Generative Adversarial Network) 모델[4]은 생성적 적대 신경망(GAN, Generative Adversarial Network)을 음성 향상에 적용한 모델로서, 생성자(Generator)와 판별자(Discriminator)라는 두 개의 네트워크가 서로 경쟁하며 학습한다. 생성자는 소음이 섞인 1차원 음성 파형을 입력받아 깨끗한 음성 파형을 생성하고, 판별자는 그 음성이 실제 깨끗한 음성인지 아니면 생성자가 만들어낸 가짜 음성인지를 판별한다. 그리고 이러한 과정의 반복적인 수행을 통해 소음 제거 성능을 개선한다.

2.1.5 소음 제거 모델 비교

과거 통계적 신호 처리 기반 소음 제거 모델들은 소음의 정상성·선형성을 가정했기 때문에, 불확실성이 강한 소음 환경에서는 성능이 심각하게 저하되는 단점이 있다. 앞서 소개한 4종의 소음 제거 모델은 각기 다른 방식으로 기존의 한계를 일부 극복하였으나, 모델별 특성에 따라 상이한 장단점이 있다.

스펙트럼 계열 중 마스킹 기반인 FullSubNet은 불필요한 성분을 억제하여 출력 안정성을 확보하기 쉽지만, 세부 정보의 복원보다는 선별에 치우치는 한계가 있다. 동일한 계열의 생성 기반인 U-Net은 복원 표현력이 커 다양한 잡음에 유연하게 대처할 수 있으나, 위상정보 손실이나 및 왜곡이 발생한다.

Wave-U-Net과 SEGAN은 1차원의 파형을 다루기 때문에 위상 손실 없이 파형 수준의 직접 복원을 지향한다. 그러나 잡음이 많은 환경에서는 스펙트럼 계열에 비해 음성에 대한 특성 분리 능력이 떨어진다. 특히 파형 직접 생성 및 적대적 학습의 특성상 학습 안정성 확보가 어렵고, 극단적 소음 입력 시 인위적 왜곡을 생성할 위험이 있다.

2.2 음성 인식(ASR) 모델

2.2.1 Whisper

OpenAI가 2022년에 공개한 Whisper 모델은 음성 인식 분야에서 기준 성능을 평가 및 비교하는 벤치마크 모델로서 널리 활용되는 모델이다[5]. 트랜스포머(Transformer) 기반의 음성 인식 모델로, 680,000 시간에 달하는 다국어 음성 데이터를 약한 지도학습(Weakly supervised learning)을 통해 학습하였다.

Whisper 모델은 파라미터 크기에 따라 여러 모델이 존재하지만, 본 연구에서는 국방 분야 특수성을 고려하여 tiny(39M 파라미터) 모델과 small(244M 파라미터) 모델을 성능 평가 대상으로 선정하였다. 실제 국방 분야에서는 엣지 디바이스(Edge device)나 온프레미스(On-premise) 환경에서 모델이 운용될 가능성이 높다. 따라서 상대적으로 경량화된 모델이 특수 환경에서도 실용적인 성능을 보일 수 있는지를 검증할 필요가 있어 두 모델을 선정하였다.

2.2.2 Wav2Vec 2.0

Meta AI에서 제안한 Wav2Vec 2.0 모델[6]은 자기 지도학습(Self-supervised learning) 기반의 모델로서, CNN 기반으로 동작했던 기존의 1.0 모델과 달리 트랜스포머 인코더를 활용한다. 레이블이 없는 대규모 음성 데이터로부터 음성의 풍부한 맥락적 표현을 학습하였으며, CTC(Connectionist Temporal Classification) Loss를 적용하여 음성 프레임과 텍스트 시퀀스를 정렬하는 것이 핵심적인 특징이다.

그러나 Wav2Vec 2.0 원본 모델은 한국어로 사전 학습 되어 있지 않기 때문에, 본 연구에서는 해당 모델을 한국어 음성 데이터세트인 Zeroth-Korea를 사용해 파인튜닝한 kresnik/wav2vec2-large-xlsr-korean 모델[7]로 실험을 진행하였다.

2.3 선행연구 고찰

S. H. Oh et al.[8]은 대학 강의에서 녹음된 일반 한국어 음성 데이터세트에 총기 소리를 무작위로 합성하여 실험에 사용할 데이터세트를 생성하였다. 그리고 U-Net 모델을 기반으로 소음을 제거한 뒤 Whisper 모델군에 입력하여 문자 오류율(CER, Character Error Rate)을 측정하였다. 그 결과, 소음 제거만으로도 small 모델의 CER이 평균 19.13%P 개선될 수 있음을 확인하였다. 소음이 제거된 데이터를 사용하여 Whisper 모델을 파인튜닝 하였을 때는 CER이 평균 40.9%P까지 개선되었다. 그러나 전장 소음이 총기 소리에만 국한되어, 실제 전장 환경에서 발생하는 다양한 소음에 대해서는 검증하지 못했다는 한계가 존재한다.

S. W. Jeong et al.[9]은 전장 소음(총성, 폭발음)을 대상으로 세 가지 소음 제거 모델(CNN, SEGAN, Wave-U-Net)의 성능을 비교하였다. 성능은 음성 품질 측정 지표인 PESQ와 STOI를 활용하여 측정되었으며, Wave-U-Net이 PESQ 1.82, STOI 0.92로 가장 우수한 음질 향상을 보였다.

S. Y. Oh et al.[10]은 음성과 비음성에 대한 엔트로피 특징 추출에 이산 퓨리에 변환(Discrete Fourier Transform, DFT)을 적용해 잡음 환경에 강인한 음성 특징을 검출하는 방법을 제안하였다. 성능은 음성 인식 성능 향상 알고리즘의 성능 검증용으로 널리 사용되는 Aurora 2.0을 사용하였으며, 그 결과 인식률이 0.22dB 개선됨을 확인하였다.

그러나 선행연구 중 두 연구[9][10]는 본 연구에서 증명하고자 하는 ASR 모델의 음성인식 성능 지표인 CER을 직접 측정하지 않고 간접적인 음질 평가지표만을 사용하여 성능을 검증하였다.

III. 연구 방법

3.1 데이터세트 생성

본 연구의 목적은 실제 전장에서 발생할 수 있는 소음 환경을 대상으로 소음 제거 모델 자체의 효율성과 ASR 모델에 적용했을 때의 성능 개선 효과를 비교 분석하는 것이다. 이러한 연구 목적의 달성을 위해 깨끗한 원본 음성 데이터와 다양한 전장 소음 데이터를 랜덤하게 합성하여 실험용 데이터세트를 구축하였다.

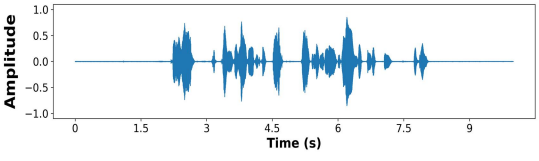
원본 음성 데이터는 2021년에 발표된 AI Hub의 '자유대화 음성(일반 남여)' 데이터세트[11]를 사용하였다. 해당 데이터세트는 조용한 실내에서 일반인 남녀가 일상적인 주제로 자유롭게 대화한 음성으로 구성되어 있으며, 2,000명 이상의 발화자를 통해서 수집된 4,000여 시간 분량의 음성 데이터로 이루어져 있다. 본 실험에서는 전체 데이터세트 중 발화의 다양성을 확보하기 위해 무작위로 선별된 22시간 분량의 음성 데이터 13,062개를 선별해 사용하였다.

전장 소음 모사를 위한 데이터세트는 군사 상황 인식 및 감시 시스템 연구를 위해 구축된 데이터세트인 MAD Dataset[12]를 사용하였다. 해당 데이터세트는 실제 군사활동을 촬영한 비디오에서 추출한 총 7개 클래스(통신, 총성, 발소리, 포격, 차량, 헬리콥터, 전투기)의 음성 데이터로 이루어져 있으며, 본 연구에서는 총성, 포격, 차량, 헬리콥터 및 전투기 5개 클래스 중 83개 소음 데이터를 사용하였다.

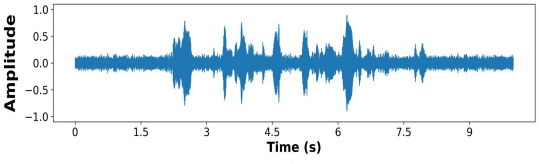
원본 및 전장 소음 데이터세트는 원본 음성 신호(Signal)와 전장 소음(Noise)의 에너지 비율인 신호대 잡음비(SNR, Signal-to-Noise Ratio)를 기준으로 무작위로 합성하였다. SNR은 다양한 소음이 발생하

는 전장환경을 모사하기 위해 '5', '0', '-10'dB의 세 가지 레벨로 설정하였다.

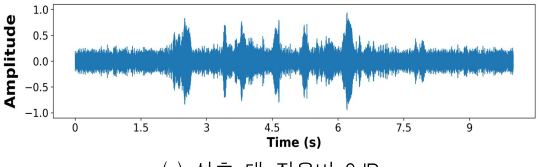
SNR 5dB는 음성 신호의 에너지가 소음보다 약 3.2배 더 강한 상황으로, 소음원이 비교적 원거리에 위치하거나 약하게 진행되는 교전 상황을 가정한다. 0dB는 음성과 소음의 에너지가 동일한 상황으로, 교전이 근거리에서 발생하여 음성 식별이 어려워진 전장환경을 가정한다. 마지막으로 -10dB는 소음 에너지가 음성 신호보다 10배 더 강한 극한의 상황으로서, 총성이나 폭음이 근처에서 발생하여 음성이 소음에 완전히 묻히는 환경을 가정한다. 원본 음성 데이터와 전장 소음 데이터가 3가지 레벨로 합성된 이후 생성된 음성 데이터의 파형은 그림 1에서 보는 바와 같다.



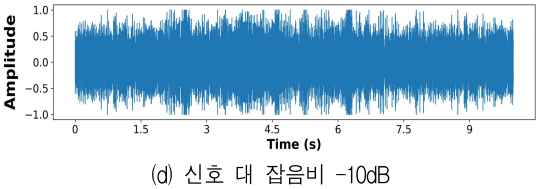
(a) 원본 무잡음 음성
(a) Original clean speech



(b) 신호 대 잡음비 5dB
(b) SNR 5dB



(c) 신호 대 잡음비 0dB
(c) SNR 0dB



(d) 신호 대 잡음비 -10dB
(d) SNR -10dB

그림 1. 다양한 수준의 음성 및 소음 데이터 합성
Fig. 1. Synthesis of data with different level

합성 데이터세트는 훈련(Train), 검증(Validation), 평가(Test) 세트로 분리되었으며, 4가지의 소음 레벨(Clean, SNR 5dB, 0dB, -10dB)에 모두 동일하게 13,062개의 데이터를 구축하였다. 합성 데이터세트 세부 구성은 표 1에서 보는 바와 같다.

표 1. 실험에 활용한 데이터세트의 구성
Table 1. Experimental dataset configuration

Dataset	Number of data	Ratio
Train	9,143	70%
Validation	1,306	10%
Test	2,613	20%
Total	13,062	100%

3.2 평가지표

ASR 모델 및 소음 제거 모델의 성능과 효율성을 종합적으로 평가하기 위한 지표로는 실시간 계수, MAC 및 문자 오류율을 사용하였다.

3.2.1 실시간 계수(RTF)

소음 제거 모델의 효율성을 측정하기 위한 속도 지표인 RTF(Real-Time Factor)는 1초 분량의 음성 데이터를 처리하는 데 소요되는 실제 시간(초)을 의미한다. 1보다 작을수록 모델의 성능이 좋으며, 음성 데이터의 실시간 처리가 가능함을 나타낸다.

3.2.2 MAC

소음 제거 모델의 연산 복잡도를 측정하기 위한 지표로, 모델이 하나의 데이터를 처리하는데 필요한 총 곱셈-누산 연산 횟수(단위 : G = 10억번)를 의미한다. MAC(Multiply-Accumulate Operations) 값이 낮을수록 모델은 더 적은 계산량으로 동작하므로 제한된 하드웨어 자원의 환경에서 유리하다.

3.2.3 문자 오류율(CER)

ASR 모델의 정확성을 측정하기 위한 핵심 성능 지표는 일반적으로 CER(Character Error Rate)을 사

용한다. CER은 ASR 모델이 예측한 스크립트와 실제 정답 스크립트 간 문자 수준 차이를 측정하는 지표이다. 계산된 값은 정답 스크립트의 총 문자 수 대비 잘못 인식된 문자의 비율을 의미하며, 계산식은 식 (1)과 같다.

$$CER(\%) = \frac{S+D+I}{N} \times 100 \quad (1)$$

S(Substitutions)는 결과에서 잘못 대체된 글자 수, D(Deletions)는 삭제된 글자 수, I(Insertions)는 잘못 삽입된 글자 수를 각각 의미하며, N은 정답 스크립트의 총 문자 수이다. CER은 값이 낮을수록 오류가 적다는 것을 의미하며, 0%에 가까워질수록 완벽한 인식 성능을 가짐을 의미한다.

3.3 실험 방법

실험은 성능 비교 대상으로 선정한 4종의 소음 제거 모델 및 3종의 ASR 모델을 대상으로 3.1절의 과정을 통해 구축된 합성 데이터세트를 적용하여 전장 소음 환경에서의 성능 변화를 단계적으로 분석한다. 세부적인 실험 과정은 총 4단계로 진행되었으며, 각 실험 단계별 세부적인 수행 내용은 다음과 같다.

3.3.1 실험 #1 : 소음 제거 모델의 효율성 비교

첫 번째 실험은 소음 제거 모델들의 효율성 비교 실험으로서, 4종의 모델별로 파라미터, 실시간 계수, MAC를 측정하여 국방환경에서의 적용 가능성을 검증한다.

3.3.2 실험 #2 : 기준 성능(Baseline) 평가

두 번째 실험은 ASR 모델의 원본 성능을 측정하는 단계이다. 3종의 ASR 모델을 4가지 환경(Clean, SNR 5dB, 0dB, -10dB)의 데이터세트를 이용해 각각 CER을 측정하고 소음 수준 변화에 따른 성능 저하폭을 확인한다.

3.3.3 실험 #3 : 소음 제거 모델 적용 평가

세 번째 실험은 음성 데이터에 소음 제거 모델만 적용했을 때의 성능을 평가하는 단계이다. 이를 위해 3가지 환경(SNR 5dB, 0dB, -10dB)의 데이터세트를 소음 제거 모델에 입력하여 소음을 제거하고, 이를 ASR 모델에 입력해 최종적으로 측정된 CER을 실험 #2의 기준 성능과 비교하여 성능 개선 효과를 평가한다. 소음 제거 모델에 적용된 하이퍼파라미터는 아래 표 2와 같다.

표 2. 소음 제거 모델에 사용된 하이퍼파라미터
Table 2. Hyperparameters used in denoising model

Denoising model	Learning rate	Batch size	Loss	Epoch
FullSubNet	1e-3	8	MSELoss	50
U-Net	1e-4	8	HuberLoss	50
Wave-U-Net	1e-4	16	Adversarial+L1	50
SEGAN	G : 1e-5 D : 1e-4	16	L1Loss	50

3.3.4 실험 #4 : 소음 제거 + ASR 모델 파인튜닝

마지막 실험은 소음 제거 전처리와 ASR 모델 파인튜닝을 결합했을 때의 성능 개선 효과를 평가하는 단계이다. 이를 위해 3가지 환경(SNR 5dB, 0dB, -10dB)의 데이터세트를 소음 제거 모델로 전처리하고, 이를 사용하여 ASR 모델을 각각 파인튜닝을 실시한 뒤 최종적으로 측정된 CER을 실험 #2 및 실험 #3의 결과와 비교를 통해 성능 개선 효과를 검증한다.

Ⅳ. 실험 결과

실험은 Window 11 Home 기반에 AMD Ryzen 7500f CPU와 NVIDIA RTX 4060Ti 16GB GPU, 32GB RAM이 탑재된 데스크탑 환경에서 진행되었으며, 사용한 개발 언어는 Python 3.12이다.

4.1 실험 #1 : 소음 제거 모델 효율성 비교 결과

표 3에서 보는 바와 같이 Wave-U-Net과 SEGAN이 다른 모델 대비 낮은 MAC과 빠른 RTF를 보였

고, U-Net의 경우 133G의 높은 MAC를 보였음에도 RTF는 0.0043으로 빨랐다. FullSubNet은 MAC은 작으나 RTF가 타 모델 대비 2배 이상 컸다. RTF를 기준으로 판단하면 모델 4종 모두 실시간 처리 능력을 가진 소음 제거 모델로 평가할 수 있다.

표 3. 실험 #1의 결과
Table 3. Result of experiment #1

Denoising model	MAC (G)	RTF	Interference time (ms)
FullSubNet	3.55	0.022	109.84
U-Net	133	0.0043	21.71
Wave-U-Net	1.53	0.0098	48.83
SEGAN	3.34	0.0031	15.53

4.2 실험 #2 : 기준 성능(Baseline) 평가 결과

표 4에서 보는 바와 같이 SNR이 5dB에서 -10dB로 낮아지면 CER이 증가하며 성능이 저하되는 현상이 공통적으로 나타났다. Whisper-tiny 모델은 0dB 이상부터는 CER이 100%를 초과하여 음성인식이 거의 불가능하였다. 또한, SNR -10dB의 극한 환경에서는 모든 모델의 CER이 80%를 초과하여 사실상 음성인식이 불가능한 수준이었다.

표 4. 실험 #2의 결과
Table 4. Result of experiment #2

SNR	Whisper tiny (%)	Whisper small (%)	Wav2vec 2.0 (%)
Clean	21.00	9.84	21.99
5dB	58.45	13.41	41.39
0dB	109.04	21.64	55.68
-10dB	706.96	237.72	82.77

4.3 실험 #3 : 소음 제거 모델 적용 평가 결과

소음 제거 모델 중 FullSubNet 모델이 다양한 SNR 레벨과 ASR 모델 조합에서 가장 뛰어난 소음 제거 성능을 달성하였다. 특히 FullSubNet과 Wav2vec 2.0 모델을 결합한 경우, SNR -10dB에서 57.78%의 가장 우수한 CER을 기록하였으며, Whisper-small과 FullSubNet을 결합했을 때에도 -10dB에서 CER이 약 157%P 개선되었다. 반면 Wave-U-Net과 SEGAN은 오히려 CER이 폭증하면서 음성인식에 실패함을 확인하였다. 세부적인 실험 결과는 표 5에서 보는 바와 같다.

표 5. 실험 #3의 결과
Table 5. Result of experiment #3

Denosing model		Whisper tiny (%)	Whisper small (%)	Wav2vec 2.0 (%)
FullSubNet	5dB	39.62	15.47	31.13
	0dB	52.10	25.74	37.16
	-10dB	128.94	80.17	57.78
U-net	5dB	59.93	16.57	34.64
	0dB	87.08	31.33	42.94
	-10dB	516.39	188.57	66.25
Wave-U-Net	5dB	144.32	134.38	85.90
	0dB	174.19	154.07	88.43
	-10dB	369.18	159.66	94.38
SEGAN	5dB	562.63	211.39	74.37
	0dB	761.93	403.37	80.42
	-10dB	692.37	404.46	93.12

4.4 실험 #4 : 소음 제거+ASR 모델 학습 결과

마지막 실험은 실험 #2와 #3에서 CER이 폭증했던 모델들을 배제하고, 유의미한 결과를 얻은 소음 제거 모델(FullSubNet, U-Net)과 ASR 모델(Whisper-small, Wav2vec 2.0)에 대해서만 진행하였다. 하이퍼파라미터는 두 모델이 동일하게 50 epochs, batch size 8, learning rate 1e-5로 설정하였으며, 실험 결과는 표 6에서 보는 바와 같다.

표 6. 실험 #4의 결과
Table 6. Result of experiment #4

Denosing model		Whisper small (Finetuned) (%)	Wav2vec 2.0 (Finetuned) (%)
FullSubNet	5dB	184.77	10.17
	0dB	184.73	14.03
	-10dB	186.82	35.36
U-net	5dB	184.76	12.48
	0dB	184.88	18.17
	-10dB	196.47	45.89

소음을 제거한 데이터로 ASR 모델을 파인튜닝한 결과, Wav2Vec 2.0 모델은 모든 소음 환경에서 가장 낮은 CER을 달성하는 동시에, 모든 소음 환경에서 CER이 20%P 이상 개선되었다. 반면 Whisper-small

모델은 모든 소음 환경에서 실험 #2 대비 CER이 180%P 이상으로 상승하며 음성인식에 실패하였다.

소음 제거 데이터로 파인튜닝을 수행했음에도 불구하고 Whisper 모델의 성능이 급격히 저하된 원인은 소음 제거 과정에서 발생한 인위적 왜곡(Artifact)이 파인튜닝을 통해 모델에 잘못 학습되었기 때문이다. Wav2Vec 2.0과 같은 인코더 기반 모델이 양방향 문맥을 고려하여 음소를 판별하는 것과 달리, Whisper 모델은 이전 시점의 출력 정보를 다시 입력으로 받아 순차적으로 텍스트를 생성한다. 따라서 파인튜닝 과정에서 왜곡된 입력 신호를 특정 단어와 잘못 매핑하게 되고, 실제 추론 단계에서 동일한 왜곡이 감지되면 디코더가 잘못된 단어를 생성하기 시작한다. 이와 같이 Whisper 모델은 이전 시점의 잘못된 출력 정보를 다시 입력으로 받아 문맥을 억지로 이어가려 하므로, 실제 음성과는 무관한 문장을 끊임없이 생성하는 환각(Hallucination) 현상이 증폭된다. 이러한 현상은 OpenAI에서도 지적한 바와 같이 입력 신호의 불확실성이 높은 구간에서 디코더가 문맥 의존성에 갇히는 현상[5]이 파인튜닝으로 인해 심화된 결과라 할 수 있다.

Ⅴ. 결론 및 향후 과제

본 연구는 음성인식 기술이 극한의 전장 소음에 노출될 경우 성능이 급격히 저하되는 문제를 해결하기 위해 적용하는 소음 제거 모델들의 효율성과 자동음성인식 모델에 대한 성능 개선 효과를 비교 평가하였다.

먼저 4종의 소음 제거 모델들(FullSubNet, U-Net, SEGAN, Wave-U-Net)을 대상으로 효율성을 측정한 후 전장 소음을 합성한 음성 데이터를 적용해 기준 성능을 측정하였다. 이어서 소음 제거 모델과 ASR 모델 3종(Whisper-tiny, Whisper-small, Wav2Vec 2.0)의 조합을 통해 음성인식 성능이 어느 정도 개선될 수 있는지를 종합적으로 비교 분석하였다.

실험 결과, 4종의 소음 제거 모델은 모두 실시간 처리에 충분한 성능을 보유하고 있음을 확인했으며, ASR 모델의 성능 개선 측면에서는 소음을 제거한 데이터로 파인튜닝하는 접근법이 모든 SNR 레벨에서 가장 우수했다. 특히 SNR -10dB의 극심한 소음 환경에서는 FullSubNet과 Wav2Vec 2.0 결합 모델이 35.83%라는 가장 낮은 CER을 달성해 소음에 가장 강건한 모델이 될 수 있음을 확인하였다.

연구의 제한사항으로는 실제 군에서 수집된 음성 데이터가 아닌 일반 대화 음성을 기반으로 소음을 합성하여 실험을 진행했기에 전장환경에의 적용을 현재 수준에서 단정할 수 없다는 점이 있다.

따라서 향후 연구에서는 실제 군 지휘통제 용어 등이 포함된 발화 데이터를 확보하거나 제작하여 연구의 실효성을 검증하고자 한다. 또한, 경량 언어 모델을 활용한 맥락화 및 도메인 특화 방법으로 음성인식 결과를 보강하여 국방 분야에 특화된 음성인식 파이프라인을 구축하는 방향으로 연구를 확장할 예정이다.

References

[1] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement", 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 6633-6637, Jun. 2021. https://doi.org/10.1109/ICASSP39728.2021.9414177.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", Medical image computing and computer - Assisted intervention - 2015: 18th international conference, Munich, Germany, pp. 234-241, Oct. 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

[3] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation", Proc. of the 19th International Society for Music Information Retrieval Conference, International Society for Music Information Retrieval, Paris, France, pp. 1-7, Sep. 2018. https://doi.org/10.48550/arXiv.1806.03185.

[4] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network", Proc. of Interspeech 2017, ISCA, Stockholm, Sweden, pp. 3642-3646, Aug. 2017. https://doi.org/10.48550/arXiv.1703.09452.

[5] A. Radford, J. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision", Proc. of the 40th International Conference on Machine Learning, Hawaii: USA, pp. 28492-28518, Jul. 2023. https://doi.org/10.48550/arXiv.1806.03185.

[6] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: a framework for self-supervised learning of speech representations", Proc. of the 34th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, No. 1044, pp. 12449-12460, Dec. 2020. https://doi.org/10.48550/arXiv.2006.11477.

[7] Wav2vec2-large-xlsr-korean, https://huggingface.co/kresnik/wav2vec2-large-xlsr-korean. [accessed: Sep. 05, 2025]

[8] S. H. Oh, J. H. Park, T. K. Yuk, J. A. Kim, and H. Kwon, "Speech recognition model robust to battlefield noise", Journal of the Korea Institute of Information and Communication Engineering, Vol. 28, No. 6, pp. 677-684, Jun. 2024. http://doi.org/10.6109/jkiice.2024.28.6.677.

[9] S. W. Jeong and J. M. Ma, "Comparison and Proposal of Denoising Models for Improving Speech Recognition Performance in Battlefield Environment", Journal of KAIS, Vol. 26, No. 2, pp. 682-688, Feb. 2025. https://doi.org/10.5762/KAIS.2025.26.2.682.

[10] S. Y. Oh, "Improvement Entropy Feature Extraction for AI Voice Recognition Improvement", Journal of KIIT, Vol. 23, No. 7, pp. 149-154, Jul. 2025. http://dx.doi.org/10.14801/jkiit.2025.23.7.149.

[11] Free Conversation Voice (General Men and Women) Dataset, https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=109. [accessed: Sep. 11, 2025]

[12] MAD Dataset: Military Audio Dataset, https://www.kaggle.com/datasets/junewookim/mad-dataset-military-audio-dataset. [accessed: Sep. 11, 2025]

저자소개

김 동 훈 (Donghoon Kim)

2017년 2월 : 육군사관학교 기계공학과(공학사)
2026년 1월 : 국방대학교 사이버·컴퓨터공학과(공학석사)
2026년 1월 ~ 현재 : 육군3사관학교 전자공학과 강사
관심분야 : 머신러닝, 정보보호, 거대언어모델, 침입탐지시스템

안 서 경 (Seokyeong An)

2017년 2월 : 성신여자대학교 법학과(법학사)
2026년 1월 : 국방대학교 사이버·컴퓨터공학과(공학석사)
2026년 1월 ~ 현재 : 육군본부 지상군페스티벌기획단 사이버대응장교
관심분야 : 음성인식, 자연어처리, 사이버전, 유·무선통신

이 수 진 (Soojin Lee)

1992년 3월 : 육군사관학교 전산학과(이학사)
1996년 2월 : 연세대학교 컴퓨터과학과(이학석사)
2006년 2월 : 한국과학기술원 전산학과(공학박사)
2006년 3월 ~ 현재 : 국방대학교 사이버·컴퓨터공학과 교수
관심분야 : 국방 사이버 보안 정책, 침입탐지시스템, 모바일 네트워크 보안, 머신러닝, 암호이론 및 응용