https://doi.org/10.33778/kcsa.2024.24.2.161

국방 분야에서 전장 소음 환경 하에 음성 인식 기술 연구★

김 영 훈*, 권 현**

요 약

최근 음성 인식 모델들이 점점 발달하고 있고 이와 더불어 좋은 데이터를 얻기 위한 다양한 음성 처리 기술들도 발전하고 있다. 한편 국방 분야에서도 노이즈가 낀 음성 데이터로부터 노이즈를 제거하고 이를 효과적으로 음성 인식하는 기술을 접목하려고 시도하고 있다. 본 논문에서는 다양한 소음이 존재하는 전장 상황 속에서 음성 인식 기술을 활용하여 효과적으로 지휘관이 명령을 전달할 수 있는 음성 인식방법을 제안하였다. 제안방법은 노이즈가 있는 음성에 대해서 노이즈를 제거 후 OpenAI의 Whisper 모델을 사용하여 텍스트로 변환하는 방법이다. 실험결과로써, 제안 방법은 노이즈를 제거하지 않은 기존 방법에 비해서 글자 오류률(Charactor Error Rate, CER)이 6.17% 감소된 것을 볼 수가 있었다. 추가적으로 제안방법을 이용하여 국방분야에 적용할 수 있는 부분에 대해서도 기술하였다.

A Study on the Effective Command Delivery of Commanders Using Speech Recognition Technology

Yeong-hoon Kim*, Hyun Kwon**

ABSTRACT

Recently, speech recognition models have been advancing, accompanied by the development of various speech processing technologies to obtain high-quality data. In the defense sector, efforts are being made to integrate technologies that effectively remove noise from speech data in noisy battlefield situations and enable efficient speech recognition. This paper proposes a method for effective speech recognition in the midst of diverse noise in a battlefield scenario, allowing commanders to convey orders. The proposed method involves noise removal from noisy speech followed by text conversion using OpenAI's Whisper model. Experimental results show that the proposed method reduces the Character Error Rate (CER) by 6.17% compared to the existing method that does not remove noise. Additionally, potential applications of the proposed method in the defense are discussed.

Key words : Deep neural network, Speech recognition model, Noise speech data

접수일(2023년 10월 01일), 게재확정일(2024년 06월 30일)

★ 본 논문은 육군사관학교 화랑대연구소의 2024년도 연구활동비 지원을 받아 연구되었음.(연구번호: 2024B1005).

* 제3군단 제20기갑여단 정보통신대 운용소대장(주저자)
** 육군사관학교 AI·데이터과학과 부교수(교신저자)

1. 서 론

국방 분야에서 4차 산업과 첨단과학기술군을 목표로 인공지능 기술을 감시체계, 무기체계 등에 접목할려고 하고 있다[1]. 그 중에서 음성인식에 대한 관심이 증대되고 있다. 전장상황에서 수많은 정보들이 무전기나 전화기를 통해서 음성으로 전달되는 상황이 있다. 하지만 이러한 음성 데이터에 노이즈가 많이 반영되면 사람이 식별하기 어려운 음성이 있고 인력부족으로 인해 많은 음성 데이터를 효과적으로 처리하는 것에 한계가 있을 수 있다. 따라서, 이러한 노이즈가 반영된 음성을 텍스트로 전환하는 기술은 국방 분야에 필요성이 있다.

음성인식 모델과 관련하여, Google이 2017년 발표한 논문 “Attention is all you need"에서 공개한 모델 트랜스포머(Transformer)의 등장 이후 음성 인식(Speech Recognition) 분야에서 다양한 기업들이 대규모 음성 데이터를 학습한 다양한 모델을 내놓고 있다[2]. ”좋은 데이터에서 좋은 결과가 나온다“라는 말이 있듯이 이러한 모델들의 등장과 더불어 좋은 데이터를 얻기 위해 다양한 음성 처리 기술들도 발전하고 있다. 좋은 음성 데이터를 구축하기 위해서, 음성 내부에 있는 노이즈를 효과적으로 제거하여 음성인식모델의 성능을 개선시키는 연구가 중요하다. 따라서 노이즈가 낀 음성 데이터에서 노이즈만 분리하여 제거하는 기술들도 다양하게 연구되고 있다.

본 연구에서 다양한 소음이 존재하는 전장 상황 속에서 음성 인식 기술을 활용하여 효과적으로 지휘관이 명령을 전달할 수 있는 음성 인식방법을 제안하였다. 다양한 소음이 존재하여 지휘관에 명령을 음성인식모델을 통해 음성 명령을 텍스트로 전환하는 것은 지휘관의 명령을 이해하는 데 도움이 될 것으로 판단하였다. 이 논문의 공헌점은 다음과 같다. 먼저, 제안된 방법에서는 음성을 텍스트로 변환하는 최신 STT 모델인 OpenAI의 whisper 모델[3]을 사용하여 노이즈가 낀 음성 데이터로부터 지휘관의 음성만을 추출하여 이를 텍스트로 변환하여 이를 수신자에게 제공하는 것에 대해서 연구하였다. 두 번째로, 음성 데이터를 텍스트 데이터로 변환하는 모델의 정확도를 높이는 방법에 대해 소개하고 이를 실험적으로 증명하였다.

이 논문의 나머지 구성은 다음과 같다. 2장에서 제안방법에 대한 관련연구를 소개하고 3장에서 제안 방법에 대해서 설명한다. 4장에서 실험 환경, 실험 결과, 분석에 대해서 기술하였다. 5장에서 국방 분야에서 적용할 수 있는 방안에 대해서 다루었고 마지막으로 6장에서 이 논문의 결론으로 구성되어 있다.

2. 관련연구

2.1 음성 데이터의 노이즈 제거 방법

전장 환경에서 발생하는 다양한 노이즈는 음성으로 전달되는 지휘관의 명령을 인식하는 데에 어려움을 초래할 수 있다. 따라서 음성 데이터에서 지휘관의 육성과 노이즈를 분리하여 노이즈를 제거하는 것은 음성 처리 분야에서 중요한 작업이다.

노이즈 제거 알고리즘은 크게 두 가지 접근 방식으로 구현된다. 먼저, 시간 도메인 노이즈 제거 알고리즘(Time-Domain Denoising Algorithm)은 음성 데이터에서 노이즈와 신호를 분리하는 과정에서 시간 도메인에서의 특성을 이용한다[4]. 일반적으로, 파동 변환과 필터링 기술을 사용하여 잡음과 음성 신호를 분리한다. 이 방법은 계산 효율이 높아서 실시간으로 응용하기에 적합하지만 시간 도메인에서 처리하기 때문에 일부 주파수 영역에서는 효과가 제한될 수 있다. 다음으로 주파수 도메인 노이즈 제거 알고리즘(Frequency-Domain Denoising Algorithm)[5]은 고속 푸리에 변환(Fast Fourier Transform)[6]를 사용하여 음성 데이터를 주파수 영역으로 변환한 후, 주파수 도메인에서의 특성을 활용한다. 주로 스펙트럼 마스킹, 위상 추정, 분해 필터링 기술을 사용하여 잡음을 제거한다. 이 방법은 주파수 도메인에서 작동하기 때문에 특정 주파수 영역에서 노이즈를 효과적으로 제거할 수 있지만 계산 비용이 높을 수 있어

리소스 소모가 크다는 단점이 있다.

2.2 Whisper 모델

2021년 12월, OpenAI는 자동 음성 인식(ASR, Automatic Speech Recognition) 모델 Whisper[3]를 공개했다. 이 모델은 다양한 언어(MultiLingual) 환경에서 제로샷(Zero-Shot)으로 뛰어난 성능을 보여 주목을 받았다. Whisper의 기초 모델은 이전에 NLP 분야에서 혁신적인 변화를 가져온 Transformer이다. Whisper Transformer의 인코더-디코더 구조를 따라 음성 데이터가 인코더에 입력되면, 해당 음성 데이터로 처리해야 하는 내용을 디코더에 입력한다. 인코더에 입력되는 음성 데이터는 30초 단위로 쪼개지며, 이후 Mel-Spectrogram을 통해 변환된다. Whipser는 여기에 로그를 취해 Log-Mel Spectrogram으로 변환하고 인코더에 입력하기 전 Convolution Layer로 임베딩한다. 디코더는 스페셜 토큰을 포함한 음성 데이터의 내용이 담기는데 스페셜 토큰은 특정 언어와 모델의 역할을 지정해주는 토큰을 의미한다. 디코더의 목표는 다음 토큰을 예측하는 것으로 학습 과정에서 인코더에서 처리한 정보를 넘겨 받게 되고, 이러한 방식으로 음성 데이터를 인식하도록 학습된다.

3. 제안방법

제안방법은 크게 2단계에 거쳐 진행되며 1단계는 노이즈 제거 단계, 2단계는 Speech-to-Text(STT)로 음성을 텍스트로 변환하는 단계이다.
1단계에서 노이즈 제거 방법는 시간 도메인 노이즈 제거 알고리즘(Time-Domain Denoising Algorithm)을 적용하였다. 왜냐하면 위 방법은 시간 도메인 노이즈 제거 알고리즘으로 지휘관의 명령이 실시간으로 전달되는 전장 상황임을 고려하여 계산 효율이 좋아 실시간으로 응용하기 적합하기 때문이다.

2단계에서 Speech-to-Text(STT) 모델을 이용하여 음성을 텍스트로 변환하는 방법이다. 모델은 Whisper 모델을 사용하였고 학습 방법은 전이학습방법을 적용하여 이미 학습이 완료된 pre-trained 모델에 학습하고자 하는 추가적인 데이터를 학습하였다. 이 모델은 한국어의 STT 기능을 지원하는 최고 성능을 보여주면서도 fine-tuning을 통해 최적화할 수 있으므로 선정하게 되었다. 먼저, 데이터 전처리 과정을 설명하면 오디오 데이터를 로드하고 리샘플링(resampling)을 실시하고 feature extractor를 통해 1차원 오디오 배열을 log-Mel spectrogram으로 변환한다. 정답 데이터인 transcript data는 tokenizer를 이용해 label ids로 변환한다. 이후 이전에 log-Mel spectrogram으로 변환한 input featur를 PyTorch tensor로 변환하고 tokenizer를 통해 변환된 label data에 패딩(padding) 작업을 거쳐 Data Collator를 선언해준다. 두 번째로 데이터 전처리가 완료된 데이터를 Whisper 모델에 추가적으로 전이학습을 함으로써, 전장상황 관련 음성에 대해서 텍스트로 전환시키는 모델을 fine-tuning 할 수가 있다.

4. 실험 및 평가

이 장에서는 제안방법을 검증하기 위하여 실험 환경 및 실험 결과에 대해서 기술하였다. 본 실험을 진행한 실험환경은 Pytorch 머신러닝 라이브러리를 사용하였으며, GPU는 NVIDIA A100을 사용하였다.

4.1 데이터셋

데이터셋은 AI 허브에서 제공되는 공개된 한국어 대화 데이터셋 [7] 및 23년도 국방 AI 경진대회에서 제공된 음성 데이터를 사용하였다. 노이즈가 낀 길이가 서로 다른 14000개의 음성 파일을 사용하였다. 이 중에 Fine-Tuning에 필요한 훈련 데이터셋으로 11200개, 검증 데이터셋으로 1400개, 테스트 데이터셋으로 1400개를 사용하였다.

4.2 분류 모델

학습 모델로는 OpenAI의 whisper-large를 사

용하였다. whisper 모델에는 tiny, base, small, medium, large가 있는데 각 모델에 따라 파라미터와 지원되는 언어가 다르지만 그 중에서 large 모델이 성능이 잘 나왔다[8]. 모델의 하이퍼 파라미터로 학습률은 1e-5이고 warmup step은 500이고 Batch size는 4로 설정하였다.

4.3 분석결과

평가지표는 단어 오류율(Word Error Rate)를 주로 사용하는 영어와 달리 교착어인 한국어의 특성으로 인해 글자 오류율Charactor Error Rate, CER)를 사용하였다.

대조군은 노이즈 제거가 진행되지 않은 데이터를 OpenAI의 Whisper에 fine-tuning 시켜 테스트를 진행하였고 실험군(제안방법)은 노이즈 제거가 진행된 데이터를 대조군과 마찬가지로 OpenAI의 Whisper에 fine-tuning 시켜 학습을 진행하였다.

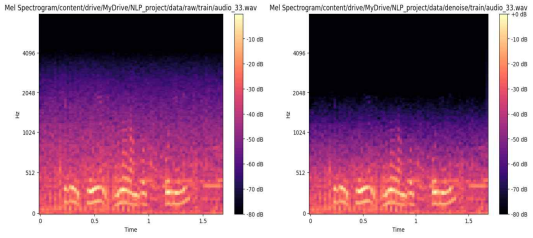


그림 1. 제안방법에 의한 노이즈 제거 전과 후 비교

실험결과 측면에서, 그림 1은 보면 제안방법에 의해서 노이즈가 제거 되기 전과 후의 스펙트럼을 보여준다. 그림 1의 오른쪽에서 1536 Hz 이상의 특정영역의 노이즈가 제거된 것을 볼 수가 있다.

표 1은 제안방법에 의해 노이즈가 제거된 전과 후의 손실함수 값과 CER 성능을 보여준다. 표 1에서 보면 노이즈 제거 후가 전에 비해 CER 6.17%p 향상하였다.

표 1. 노이즈 제거 전후 손실함수 값과 CER 비교

	손실함수 값	CER
노이즈 제거 전 (기존 방법)	0.2979	32.16%
노이즈 제거 후 (제안 방법)	0.3070	25.99%

5. 국방 분야에서의 제안방법 활용

공헌점 측면에서, 전장소음 등이 있는 상황에서 음성을 텍스트로 변환시키는 기술이다. 지휘소에서 무전기, 전화기 등을 통하여 음성을 수신하게 된다. 하지만 전장소음으로 해당 음성에 대해서 정확히 식별하기 제한되는 상황이 있고 감시체계 등으로부터 적부대 위치 정보, 규모, 부대특성 등에 대한 중요정보를 정확히 전달할 필요가 있다. 따라서 노이즈가 있는 음성을 텍스트로 전환하는 기술은 국방분야에서 중요하다. 그러한 점에서 이 연구는 노이즈가 있는 음성 데이터에 대해서 실험적으로 제안방법의 가능성과 방법론을 제시했다는 점에서 의미가 있다고 본다.

연구의 전제조건 및 한계점 측면에서, 제안 방법은 노이즈가 반영된 공개된 음성데이터셋을 사용하였다. 공개된 음성 데이터셋의 경우, 음성의 노이즈를 가우시안 노이즈를 사용하였다. 하지만 실제 전장환경에서 소음은 가우시안 노이즈가 아닌 것이기 때문에 실제 환경에 맞는 노이즈 음성 데이터셋 구축이 필요하다. 따라서 차후 실제 전장소음이 반영된 음성 데이터셋에 대한 데이터 전처리 방법에 대해 개선 할 필요가 있다.

군사적 활용 측면에서, 국방부에 있는 데이터 중에 음성 데이터에 노이즈가 반영된 것을 제거하는 기술도 중요하지만 음성 데이터에서 특정 부분이 제대로 들리지 않았지만 그것에 대해서 분류하거나 예측하는 모델에 대한 필요성이 있을 수 있다. 따라서 향후 연구에는 누락된 음성 구간을 예측하는 연구로 확장할 수가 있다.

6. 결 론

본 논문에서는 다양한 소음이 존재하는 전장 상황 속에서 음성 인식 기술을 활용하여 효과적으로

지휘관이 명령을 전달할 수 있는 음성 인식방법을 제안하였다. 제안한 방법은 노이즈가 낀 음성 데이터로부터 시간 도메인 제거 알고리즘을 사용하여 노이즈를 제거한 후 Whisper모델을 사용하여 텍스트로 변환하는 방법을 적용하였다. 실험환경으로 AI 허브에서 제공되는 한국어 대화 데이터셋과 2023년도 국방 AI 경진대회 음성데이터셋을 활용하였고 whisper 모델을 이용하였다. 실험결과로써 노이즈 제거 후 인식률은 노이즈 제거 이전 인식률에 비해 CER 6.17%p 향상을 보였다.

향후 연구로 노이즈 제거 측면에서 딥러닝 모델을 활용할 수가 있다. 예를 들어, 디노이징 오토인코더를 통하여 음성 부분에 노이즈를 효과적으로 제거하는 것이 향후연구에 가능하다. 또한, 데이터셋 구축측면에서 실제 군사작전에서 사용되는 음성데이터 구축이 향후 연구 주제가 될 수 있다. KCTC 훈련 등을 통해서 군사작전 관련 음성 데이터를 구축하고 이에 대한 데이터 라벨링 등 작업을 할 경우, 보다 실제적인 데이터셋 구축으로 의미있는 향후 연구 주제가 될 것이다.

참고문헌

[1] 조동연. "첨단기술 발전과 미래전 양상 변화에 따른 군 핵심역량 발전 방향 제시." 국방과 기술 510 (2021): 70-81.

[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[3] Amorese, Terry, et al. "Automatic speech recognition (ASR) with Whisper: Testing Performances in Different Languages." (2023).

[4] Ljubenović, Marina, et al. "Joint deblurring and denoising of THz time-domain images." IEEE Access 9 (2020): 162-176.

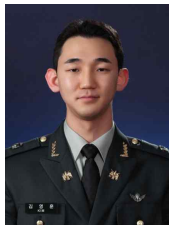
[5] Veeraiyan, Vijayabaskar, Rajendran Velayutham, and Mathews M. Philip. "Frequency domain based approach for denoising of underwater acoustic signal using EMD." Journal of Intelligent Systems 22.1 (2013): 67-80.

[6] Brigham, E. Oran, and R. E. Morrow. "The fast Fourier transform." IEEE spectrum 4.12 (1967): 63-70.

[7] https://aihub.or.kr.

[8] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever. “Robust Speech Recognition via Large-Scale Weak Supervision”.

〔저자소개〕



김 영 훈 (Younghoon Kim)
2024년 2월 육군사관학교 이학사
email : kyhkyhkyh0903@gmail.com

권 현 (Hyun Kwon)
2010년 2월 육군사관학교 이학사
2015년 8월 KAIST 전산학부 공학석사
2020년 2월 KAIST 전산학부 공학박사
email : hkwon.cs@gmail.com