Journal of the Korea Academia-Industrial cooperation Society
Vol. 25, No. 3 pp. 102-107, 2024

https://doi.org/10.5762/KAIS.2024.25.3.102
ISSN 1975-4701 / eISSN 2288-4688

트랜스포머 기반 군사용 자동음성인식 모델 제안 및 평가

차현우, 마정목*
국방대학교 국방과학학과

Proposal and Evaluation of Military Automatic Speech Recognition Model based on Transformer

Hyen Woo Cha, Jung Mok Ma*
Department of Defense Science, Korea National Defense University

요 약 현대 군사 환경에서는 정확하고 신속한 통신과 정보전달의 중요성이 강조되고 있다. 특히 전투 현장에서 양손에 무기를 들고 전투를 실시하는 군인에게 있어 음성인식 기반의 정보전달 및 무인 무기체계 조종은 필수적이다. 본 연구에서는 트랜스포머 기반 음성인식 모델을 한국어 군사용어 음성 데이터를 군사 교범, 군사용어사전, 군사 간행물로 분류하여 각각 파인튜닝하여 성능을 평가하고 비교분석 하였다. 파인튜닝한 3개의 모델 모두 기존 음성인식 모델보다 나은 성능을 보였으며, 특히 군사 교범을 학습한 모델이 가장 성능이 뛰어났다. 이는 같은 군사 데이터라도 군 내에서 사용되는 전술적 개념이나 특징 등이 포함된 데이터를 학습하는 것이 음성인식 모델의 성능을 더 많이 향상시킨다는 것을 알 수 있었으며 이를 통해 트랜스포머 기반 음성인식 모델의 군사적 적용 가능성을 확인할 수 있었다.

Abstract In contemporary military settings, emphasis is placed on the importance of precise and swift communication and dissemination of information. The utilization of speech recognition for information relaying and the operation of unmanned weapon systems is crucial, especially for soldiers involved in combat who are required to carry weapons with both hands. In this investigation, speech recognition models based on the Transformers model were fine-tuned using Korean military terminology speech data, which was categorized into military regulations, military glossaries, and military publications. The performances of these three fine-tuned models were assessed and compared. All three performed better than existing speech recognition models, and the model trained on military regulations performed exceptionally well. The study suggests that incorporating tactical concepts and characteristics utilized in the military into the data learning process can markedly enhance the performance of speech recognition models when applied to military data. The study validates the potential military applications of Transformer-based speech recognition models.

Keywords : Automatic-Speech-Recognition, Transformer, Future-Warfare, End-to-end, Whisper

1. 서론

현대 군사 환경은 다차원 전쟁(multi domain warfare)으로서 지상 ,해상, 공중, 우주, 전자기, 사이버, 인지영역 등 여러 차원에서 시간과 공간의 제약을 뛰어넘어 작전이 이루어지고 있다. 이러한 다차원에서의 작전이 이루어지기 위해선 수많은 정보가 동시에, 막힘없이 유통되어야 한다[1]. 군사 환경에서 유통되는 수많은 형태의

*Corresponding Author : Jung Mok Ma(Korea National Defense Univ.)
email: jxm1023@gmail.com
Received December 18, 2023 Revised January 12, 2024
Accepted March 8, 2024 Published March 31, 2024

정보 중 음성 신호는 단순히 주파수 변환만 실시하여 전달하기에는 다양한 환경에서 발생하는 배경 소음, 특수한 억양, 압축된 통신 등의 요소로 인해 전통적인 음성인식 기술의 한계가 존재한다. 이러한 문제점을 해결하기 위해서 자동음성인식(ASR: Automatic Speech Recognition) 기술을 적용할 수 있다. 이는 음성을 텍스트로 변환하는 기술이며 전통적인 통계 기반 시스템[2]으로부터 최근 딥러닝 기반의 종단형(end-to-end) 모델로 지속 발전되고 있다.

기존의 음성인식 기술은 은닉 마코프 모델(HMM: Hidden Markov Model)로 통계 기반의 음성인식 기술이었으며, 음향모델(AM: Aucoustic Model)과 언어 모델(LM: Language Model)을 결합하여 음성인식을 실시하였다. 최근에는 딥러닝 기반 기술을 적용하여 음성 신호의 언어적 특징 패턴을 분류하여 인공신경망으로 학습한다. 트랜스포머(Transformer) 모델은 종단형 음성인식 기술로서 대규모 데이터를 사전 학습한 모델을 미세조정하여 원하는 역할을 수행하는 모델을 만들 수 있다[3-5]. OpenAI에서는 위스퍼(Whisper) 모델을 공개하여 뛰어난 음성인식 성능을 보였다[6]. 위스퍼 모델은 잡음 환경이나 비정형 자유발화 등 기존 음성인식 모델이 어려움을 겪는 영역에서 잘 동작하는 장점이 있고, 다국어 음성인식도 가능하다. 하지만 언어마다 인식률이 다르고, 일부 언어는 단어 오류율(WER: Word Error Rate)이 30%를 넘는다. 한국어는 발음 규칙이 복잡하여 음성인식에 더욱 어려움이 있으며, 특히 군사용어는 일반적으로 사용하지 않는 특수 용어들이 다수 존재하여 추가적인 학습을 통해 성능 개선이 필요하다.

본 연구에서는 한국어 군사용어 데이터 세트를 군사교범, 군사용어사전, 군사 간행물로 구분하여 위스퍼 모델을 학습하여 한국어 음성인식 성능을 개선한다. 따라서 트랜스포머 기반 군사용 자동음성인식 모델을 제안하고 군사용어 데이터로 학습, 기존 모델과의 비교를 통해 성능을 평가하여 제안하는 모델의 군사적 활용 가능성을 확인해보았다. 이를 통해 군사 음성통신의 속도와 정확성을 높일 수 있을 것으로 기대된다.

제2장에서는 관련 연구로 트랜스포머 기반 음성인식 기술과 위스퍼 모델에 대해 설명하고, 제3장은 군사용어 데이터를 학습한 군사용 자동음성인식 모델을 소개한다. 또한 제4장은 실험 및 결과 분석을 통해 모델의 성능을 평가하고 분석하며, 제5장은 연구를 통해 얻은 시사점과 발전 방향에 대해 논의하고자 한다.

2. 관련 연구

2.1 트랜스포머 기반 자동음성인식 기술

Vaswani et al.은 입력된 시퀀스의 중요한 부분에 초점을 맞추어 모델이 문맥을 더 잘 이해할 수 있게 하는 어텐션(attention) 기법 기반의 기계 번역 모델인 트랜스포머 모델을 제안했다[7]. 이 모델은 기존의 기계 번역 분야에서 가장 우수한 성능을 보인 순환 신경망(RNN: Recurrent Neural Network) 기반 딥러닝 모델보다 더 좋은 번역 성능을 보였다. 트랜스포머 모델의 가장 큰 특징 중 하나는 병렬 처리를 통해 데이터 처리 과정을 가속화한 점이다. 이는 순환신경망 기반 모델에서 발생하는 순차적인 데이터 처리 문제를 해결했다는 것이다. 트랜스포머 모델은 인코더-디코더 구조를 가지며 시퀀스로 구성된 다양한 입출력 데이터에 대해 종단형 학습 기법을 적용한다. 기계 번역 뿐만 아니라 다양한 음성 인식을 비롯한 자연어처리(NLP: Natural Language Process) 분야에 활용되어 우수한 성능을 보여주고 있다.

트랜스포머 기술을 적용한 대표적인 음성인식 모델은 Wav2vec2.0 모델이 있다[8]. 이 모델은 facebook AI에서 발표한 음성인식 모델로 53,000 시간의 라벨링 없는 대량의 데이터로 자기지도학습을 한 모델을 10분 정도의 라벨링 된 소량의 데이터로 미세조정하여 높은 성능의 음성인식 성능을 보여줬다. Wav2vec2.0 모델은 기존 Wav2vec 모델에 비해 트랜스포머를 적용하여 성능을 더 높였으며 음성 데이터10분(평균 12.5초, 40문장)으로 미세조정 했을 때, WER 5.2%를 얻었다.

2.2 위스퍼 모델

Alec Radford at al.은 기존 트랜스포머 기반 음성인식 모델인 Wav2vec2.0 모델의 단점을 보완하여 위스퍼 모델을 제안하였다[6]. 이 모델은 680,000시간을 학습한 다국어 음성인식 모델이며 트랜스포머의 인코더-디코더 구조를 기본으로 약한 지도학습을 통해 사전 훈련하여 더욱 강건한 성능을 얻었다. 위스퍼 모델은 다국어 음성인식과 음성번역, 언어 식별까지 수행할 수 있으며, 주변 소음, 잡음 등 음성인식이 어려운 환경에서도 언어 인식 및 번역 기능이 우수하다는 장점이 있다.

위스퍼 모델의 음성인식 학습 방식은 Fig. 1과 같다. 음성 훈련 데이터(680,000시간)는 16,000Hz로 리샘플링 되며 10ms간격의 80채널의 로그-멜 스펙트로그램 형태로 모델에 입력된다. 이후 2개의 컨볼루션 레이어와

GELU 함수를 거쳐 처리된 후 토큰화 되어 여러개의 트랜스포머의 인코더와 디코더 블록을 거치며 어텐션 계산이 수행되어 최종 예측 토큰이 출력된다. 이 토큰은 다중 언어 음성 인식, 음성 번역, 말의 언어 식별 및 음성 활동 감지를 포함한 여러 음성 처리 작업에 공통으로 적용 가능하여 위스퍼 모델을 통해 음성 처리 관련 멀티태스크를 가능하게 한다.

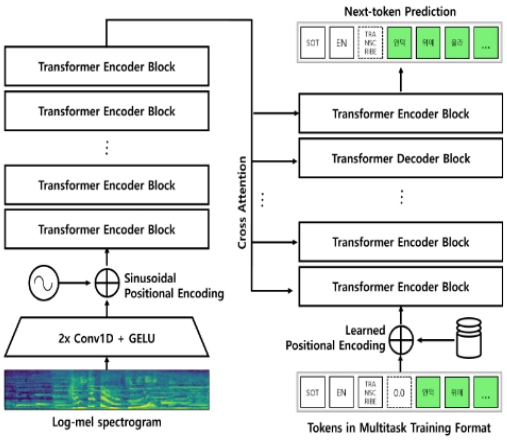

Fig. 1. Sequence-to-sequence learning in whisper (revised from [2])

3. 군사용 자동음성인식 모델

본 논문에서는 트랜스포머 기반 음성인식 모델(위스퍼)을 바탕으로 한국어 군사용어 데이터를 문서별로 나누어(군사 교범, 군사용어사전, 군사 간행물) 파인튜닝한 모델을 제안하고, 기존 트랜스포머 기반 음성인식 모델인 파인튜닝하지 않은 기존 위스퍼 모델과의 성능을 비교하여 제안하는 모델의 효용성을 판단한다.

3.1 학습 방식

Fig. 2는 군에서 구축한 음성 데이터를 위스퍼 모델에 파인튜닝하여 학습하는 과정을 나타낸다. 데이터는 군사 교범, 군사용어사전, 군사 간행물 등의 문서를 현역 남군 10명(육군 소령 2명, 육군 대위 3명, 해군 대위 2명, 공군 대위 2명, 해병 대위 1명)이 녹음을 실시한 데이터로 본 연구에서는 군사 교범, 군사용어사전, 군사간행물 데이터를 각각 10시간씩 학습한 모델을 만들어 성능평가를 실시하였다.

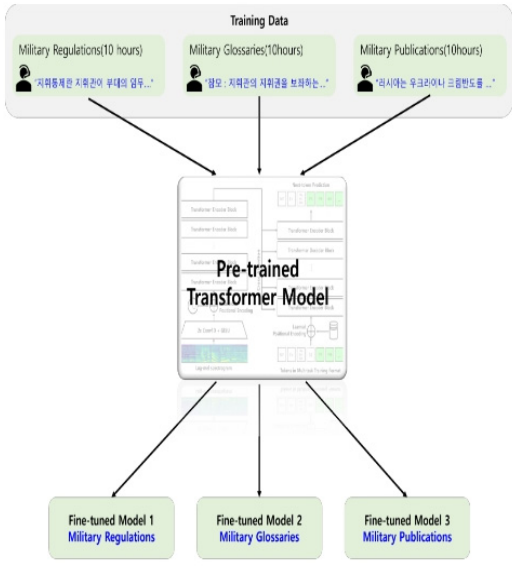

Fig. 2. Fine-tuning of pre-trained model

학습이 완료된 3가지 모델은 학습에 사용되지 않은 평가 데이터를 통해 성능을 평가하게 되고, 학습하지 않은 기존 위스퍼 모델 역시 평가 데이터를 통해 성능을 평가하여 제안하는 모델과의 비교를 실시하였다.

4. 실험 및 결과 분석

4.1 실험 구성

실험은 국방 표준 자연어 데이터셋 구축사업 분석용 PC(GPU : GeForce RTX 3090) 상에 트랜스포머 기반 음성인식 모델을 구현한 뒤 데이터 학습 및 모델 성능 평가를 실시하였다.

Table 1. Datasets used for training & evaluation of models

Data		Time(hours)
Train	Military Regulations	10
	Military Glossaries	10
	Military Publications	10
Test	Mix	2
All		32

파인튜닝을 위한 데이터 셋은 국방 표준 자연어 데이터셋 구축 사업을 통해 구축된 군사 교범, 군사용어사전,

군사 간행물에 대한 녹음파일 각 10시간이며, 평가 데이터는 학습에 사용되지 않은 음성파일 2시간을 사용했다. 모델별 학습 시간은 평균 0.25시간이 소요되었다.

실험에 적용한 하이퍼파라미터는 Table 2와 같다. 아래의 값들은 학습 데이터 셋을 8:2로 학습 데이터와 검증 데이터로 나누어 5겹 교차검증을 통해 설정하였다[9].

Table 2. Hyper-parameters applied to experiments

Hyper-parameters	Value
Learning-rate	1e-05
Batch size	32
Number of Epochs	10
Weight Decay	0.01
Optimizer	Adam

4.2 모델 성능평가 지표

제안한 모델의 성능을 평가하기 위해 훈련 데이터와 테스트 데이터에 대한 문자 오류율(CER: Character Error Rate)과 단어 오류율(WER: Word Error Rate)을 사용하여 평가하였다. 단어 오류율은 단어 단위로 오류를 계산하는 반면 문자 오류율은 음성 토큰(음절 또는 문자) 단위로 오류를 계산하여 모델의 성능을 더 세부적으로 평가하고 분석할 때 사용한다. 따라서 문자 오류율이 띄어쓰기가 잘 지켜지지 않는 한국어 음성인식의 성능을 평가하는데 더 유효하다고 볼 수 있다[10].

$$CER = (S_c + D_c + I_c)/N_c \quad (1)$$

Where, S_c denotes substitution, D_c denotes deletion, I_c denotes insertion, N_c denotes total numbers

(1)에서 S_c는 정답에는 없지만 모델이 예측한 문자의 수, D_c는 정답에는 있지만 모델 예측에서 빠진 문자의 수, I_c는 모델이 정답과 다르게 예측한 문자의 수, N_c는 정답에 있는 전체 문자의 수이다. 문자 오류율을 통해 문자 수준에서 얼마나 정확하게 예측하였는지 알 수 있다.

$$WER = (S_w + D_w + I_w)/N_w \quad (2)$$

Where, S_w denotes substitution, D_w denotes deletion, I_w denotes insertion, N_w denotes total numbers

(2)의 경우, (1)에서 글자 수였던 기준을 단어 수를 기준으로 계산하는 방식으로 단어 수준에서 얼마나 정확하게 예측되었는지 알 수 있는 지표이다.

4.3 모델별 성능 비교

학습이 완료된 각각의 모델과 학습하지 않은 기존 위스퍼 모델에 평가 데이터를 적용하여 성능평가를 실시하였으며 결과는 Table 3과 같다.

Table 3. Performance evaluation result by medels

Model	CER(%)	WER(%)
Military Criminal	7.94	4.21
Military Dictionary	12.36	6.97
Military Publication	8.42	4.54
Existing Whisper	13.68	7.01

실험 결과를 살펴보면 전체적으로 단어 오류율이 문자 오류율 보다 값이 낮은 것으로 나타났다. 이는 단어 기준으로 성능을 평가하였을 때가 문자를 기준으로 성능을 평가하였을 때와 차이가 있음을 알 수 있다. 한국어 음성인식을 성능을 평가하는데 더 유효한 문자 오류율을 모델별로 비교해 보면, 군사 교범을 학습한 모델이 7.94%, 군사용어사전을 학습한 모델이 12.36%, 군사 간행물을 학습한 모델이 8.42%로 세 모델 모두 기존 위스퍼 모델(13.68%)보다 좋은 성능을 보임을 알 수 있다. 또한 세 모델 중 군사 교범을 학습한 모델이 가장 성능이 좋으며 군사용어사전을 학습한 모델이 성능이 가장 낮음을 알 수 있었다. 이는 군사용어사전은 단순 군사용어에 대한 설명일뿐 전술적 개념이나 군 내에서만 사용되는 표현들이 없어 기존 위스퍼 모델과 비슷한 성능을 보인 것으로 판단된다. 반면에 군사교범과 군사간행물을 학습한 모델은 군사용어사전의 내용도 포함하면서 이와 관련된 전술적 내용, 시사점 등 더 다양한 문장들이 포함되어 있어서 평가 데이터에 대한 성능이 좋게 나온 것으로 판단된다.

5. 결론

본 연구에서는 군사용 음성인식 분야에서의 성능 향상을 목표로 트랜스포머 기반 음성인식 모델을 제안하고 평가하였다. Open AI에서 발표한 최신의 음성인식 모델인 위스퍼 모델을 기반으로 군에서 구축한 음성 데이터

를 군사 교범, 군사용어사전, 군사 간행물로 구분하여 각각을 학습시킨 모델과 학습을 하지 않은 기존의 위스퍼 모델의 성능을 평가하고 비교하였다. 그 결과 파인튜닝한 세 모델 모두 군사 데이터에 대해 기존 모델에 비해 향상된 성능을 보임을 확인하였고, 특히 군사 교범을 학습한 모델이 가장 우수한 성능을 보임을 알 수 있었다. 이는 단순히 군사용어만을 학습한 모델보다 군 내에서 사용되는 전술적 개념, 표현 등을 추가로 학습한 모델이 음성인식 성능이 뛰어나며 제안하는 모델의 군 적용 가능성을 확인할 수 있었다.

다만, 데이터의 균형을 위해 데이터의 양이 가장 적은 군사용어사전 데이터에 맞추어 군사 교범, 군사 간행물 데이터도 10시간으로 맞추어 학습을 하여 상대적으로 학습할 수 있는 데이터가 적었다는 제한사항과 보안상의 문제로 더 다양한 실험을 할 수 없었다는 제한사항이 있었다. 이는 앞으로 군에서 구축하는 데이터가 많아지고 다양해지고 있으며 이를 활용하기 위한 활동들이 많아지고 있으므로 가까운 시일 내에 극복할 수 있을 것으로 기대된다.

본 연구에서는 대량의 음성 데이터로 사전학습된 음성인식 모델인 위스퍼 모델을 파인튜닝한 모델을 제안하였으며 이 모델을 활용하여 AI기술이 접목된 무인 무기체계를 음성으로 통제하는 등의 군사 분야에서 활용될 수 있을 것으로 기대된다. 더 나아가, 소음제어, 운용 중인 무기체계와의 연동, 군사 네트워크 연결 등 실제 군사 환경에서의 적용을 위해 군에서 활용하는 여러 모델을 음성인식과 결합하여 미래 군 작전에 적용하기 위한 추가적인 실험과 연구를 통해 모델을 더욱 발전시키고 확장하는데 도움을 줄 수 있을 것으로 기대된다.

References

[1] K. Y. Chung, "The Fourth Industrial Revolution and the US Initiative on the Future Warfare: Analyzing the Role of Artificial Intelligence and Autonomous Weapon System", *Journal of International Politics*, Vol27, No.1, pp.5-36, Jun, 2022.
DOI: https://doi.org/10.18031/jip.2022.6.27.1.5

[2] Mark Gales, Steve Young, "The Application of Hidden Markov Models in Speech Recognition, *the essence of knowledge(NOW)*, Vol.1, No.3, pp.195-304, 2008.
DOI: https://doi.org/10.1561/2000000004

[3] W. Chan, N. Jaitly, Q. Le and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, pp. 4960-4964, 2016.
DOI: https://doi.org/10.1109/ICASSP.2016.7472621

[4] Y. M. Park, C. Y. Kim, "A Study on the Application of Language Model to Improve Speech Recognition Accuracy", *Korea Software Congress 2021*, KIIS, Pyeongchang, KOREA, pp.287-289, Dec, 2021.

[5] K. H. Shim, *Efficient Transformer Network-based End-to-End Speech Recognition.*, Ph.D's thesis, Seoul National University of Computer Engineering, Seoul, Korea, pp.14-16, 2022.

[6] Alec Radford, J. W. Kim, Tao Xu, Greg Brockman, Christine Mcleavey, Ilya Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision", *Proceedings of the 40th International Conference on Machine Learning*, pp.28492-28518, Honolulu, HI, Dec, 2022.
DOI: https://doi.org/10.48550/arXiv.2212.04356

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is AllYou Need," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, Jun, 2017.
DOI: https://doi.org/10.48550/arXiv.1706.03762

[8] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", *Proceedings of theAdvances in Neural Information Processing Systems*, pp.12449-12460, Jun, 2020.
DOI: https://doi.org/10.48550/arXiv.2006.11477

[9] J. H. Won, J. M. Shin, J. H. Kim, & J. W. Lee, "A Survey on Hyperparameter Optimization in Machine Learning", *The Journal of Korean Institute of Communications and Information Sciences*, Vol.48 No.6, pp.733-747, Jun, 2023.
DOI: https://doi.org/10.7840/kics.2023.48.6.733

[10] C. H. Oh, C. B. Kim, K. Y. Park, "Building robust Korean speech recognition model by fine-tuning large pretrained model", *Phonetics and Speech Sciences*, Vol.15, No.3, pp. 75-82, Sep, 2023.
DOI: https://doi.org/10.13064/KSSS.2023.15.3.075

차 현 우(Hyen Woo Cha) [준회원]

- 2017년 2월 : 육군사관학교 정보과학과 (정보과학 학사)
- 2024년 1월 : 국방대학교 관리대학원 국방과학학과 (무기체계 석사)

〈관심분야〉
미래전, 인공지능 참모, 자연어 처리

마 정 목(Jung Mok Ma) [정회원]

- 2002년 2월 : 육군사관학교 운영분석학과 (운영분석 학사)
- 2008년 8월 : 미국 펜실베니아주립대(PSU) (산업공학 석사)
- 2015년 5월 : 미국 일리노이대(UIUC) (산업공학 박사)
- 2015년 9월 ~ 현재 : 국방대학교 국방과학과 교수

〈관심분야〉
국방 모델링 및 데이터 분석학, 무기체계 획득관리