pISSN 2005-8063
eISSN 2586-5854
2023. 9. 30.
Vol.15 No.3
pp. 75-82

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2023.15.3.075

Building robust Korean speech recognition model by fine-tuning large pretrained model*

Changhan Oh · Cheongbin Kim · Kiyoung Park**

Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea

Abstract

Automatic speech recognition (ASR) has been revolutionized with deep learning-based approaches, among which self-supervised learning methods have proven to be particularly effective. In this study, we aim to enhance the performance of OpenAI's Whisper model, a multilingual ASR system on the Korean language. Whisper was pretrained on a large corpus (around 680,000 hours) of web speech data and has demonstrated strong recognition performance for major languages. However, it faces challenges in recognizing languages such as Korean, which is not major language while training. We address this issue by fine-tuning the Whisper model with an additional dataset comprising about 1,000 hours of Korean speech. We also compare its performance against a Transformer model that was trained from scratch using the same dataset. Our results indicate that fine-tuning the Whisper model significantly improved its Korean speech recognition capabilities in terms of character error rate (CER). Specifically, the performance improved with increasing model size. However, the Whisper model's performance on English deteriorated post fine-tuning, emphasizing the need for further research to develop robust multilingual models. Our study demonstrates the potential of utilizing a fine-tuned Whisper model for Korean ASR applications. Future work will focus on multilingual recognition and optimization for real-time inference.

Keywords: deep learning, machine learning, speech recognition

1. 서론

자동음성인식(automatic speech recognition, ASR 또는 speech to text, STT) 기술은 기계를 통해 인간의 음성을 텍스트로 변환하는 기술이다. 최근에 음성인식 기술은 종단형(end-to-end) 모델에 초점을 맞춘 딥러닝 기반 접근 방식의 등장으로 인해 전통적인 통계 기반 시스템으로부터 다양한 기술적 변화를 겪고 있다. 음성인식에서의 딥러닝 기반 접근 방식은 과거 은닉 마코프 모델(hidden Markov model, HMM)을 기반으로 하는 통계 모델과는 달리 음성 신호와 언어적 특징 간의 패턴 분류를 인공신경망

* This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2022-0-00989, Development of Artificial Intelligence Technology for Multi-speaker Dialog Modeling).

** pkyoung@etri.re.kr, Corresponding author

Received 11 August 2023; Revised 14 September 2023; Accepted 14 September 2023

을 통해 학습한다. 또한, 기존에는 음향 모델(acoustic model, AM), 언어 모델(language model, LM), 발음 사전(graphme-to-phoneme) 등의 여러 모듈을 별도로 개발하여 사용했다. 그러나 종단형 음성인식은 단일 모듈로 최종 출력단의 손실함수만으로 전체 네트워크를 학습하는 방식을 통해 모델에 입력된 음성과 정답 라벨만으로 음성인식 작업을 학습할 수 있다(Chan et al., 2016; Graves & Jaitly, 2014; Graves et al., 2006; Watanabe et al., 2018).

최근 음성인식을 위한 딥러닝 모델은 Transformer 모델을 기본으로 그 구조와 성능의 발전을 거듭하고 있다(Vaswani et al., 2017). 또한 모델의 구조가 복잡해지고 크기가 커짐에 따라 대규모 데이터를 이용하여 사전학습된 모델을 이용하여 각 응용분야에 맞는 모델을 파인튜닝하는 학습방법이 일반화되었다. 사전학습 모델을 학습하는 방법으로는 비전사 데이터를 대규모로 수집하여 자체 지도학습(self-supervised learning, SSL)을 수행하는 방법이 널리 활용되고 있다. SSL 기법은 대표적으로 자연어처리 분야에서 구글의 BERT를 통해 증명되었으며, 다양한 벤치마크에서 최고 수준의 성능을 달성하였다. 이에 따라 자연어처리 분야 뿐만 아니라 이미지, 음성 등 여러 분야에서 SSL 모델들이 등장했다(Devlin et al., 2018).

SSL 모델은 주로 자가 예측(self-prediction) 또는 대조 손실 학습 등을 진행할 수 있는 레이블과 목적 함수를 설정한다(Hadsell et al., 2006). 또한 Transformer 인코더 구조를 기반으로 하여 대량의 데이터를 학습할 수 있다. 대량의 레이블 없는 데이터로 사전 학습한 모델은 음성인식 작업과 그 외 다양한 음성 기반 작업에서 최고수준의 성능을 달성하거나 기존 기법에 준하는 결과를 기록했다(Chen et al., 2020; van den Oord et al., 2018). 대표적인 모델로 wav2vec 2.0과 Hubert 등의 모델이 있으며, 이렇게 사전 훈련된 모델은 매주 작은 규모의 음성 데이터로 파인튜닝된 경우에도 효과적인 성능을 보여주었다(Baevski et al., 2020; Schneider et al., 2019; Tsai et al., 2019).

음성인식 분야에서의 SSL 모델들이 발전을 거듭하는 가운데 최근 OpenAI에서는 Whisper 모델을 공개하여 뛰어난 성능을 보여주었다(Radford et al., 2023). Whisper 모델은 비전사데이터를 사용하는 사전학습 모델과는 달리 대규모의 데이터를 인터넷을 통하여 수집하고, 이를 자동으로 정제하여 대규모의 전사데이터를 만들어 학습하는 방식을 적용하였다. 대규모의 데이터를 사용하여 훈련함으로써 Whisper 모델은 파인튜닝 없이도 우수한 음성인식 성능을 보이며, 특히 잡음 환경이나 비정형 자유발화 등 기존 음성인식 모델이 어려움을 겪는 영역에서 잘 동작하는 장점을 가지고 있다. 또한 96개의 언어에 대한 음성 인식 기능을 제공하여 다국어 인식이 필요한 다양한 분야에 활용될 수 있다. 하지만 다국어 음성인식 모델의 고질적인 문제인 데이터양과 질의 문제는 강건한 다국어 음성인식의 성능을 저해하는 요인이 된다. Whisper 모델의 경우에도 언어마다 인식률이 다르며, 몇몇 언어에서는 단어 오류율(word error rate, WER)이 30%를 초과한다. Whisper 모델의 확장성과 성능 안정성을 보장하기 위해서는 인식 오류율이 높은 언어에 대해 성능 개선이 필요하다.

한국어는 특히 발음 규칙의 복잡성으로 인해 음성인식에 더욱 어려움이 있을 수 있다(Nam, 2019). 본 연구에서는 Whisper 모델의 한국어 성능 향상을 목표로 하여 실험을 수행한다. 이를 통해 인식 오류율이 높은 언어에서의 Whisper 모델의 성능 향상 가능성 확인과 검증에 초점을 두고 연구를 수행한다. 연구에서는 추가적인 한국어 데이터 세트를 통해 Whisper 모델을 파인튜닝하여 한국어 음성인식을 수행하여 한국어 음성인식 성능을 개선한다. 이 개선된 성능을 사전학습 모델을 이용하지 않고 대용량 데이터를 사용하여 학습된 Transformer 모델의 성능과 비교한다.

본 연구의 기여는 다음과 같다. 첫 번째로, 우리는 한국어 데이터에 대한 Whisper 모델의 성능 향상을 위한 실용적 방법을 제안하고, 실험을 통해 성능을 평가했다. 두 번째로, Whisper 모델은 처음에는 파인튜닝 없이 바로 사용할 수 있도록 설계되었지만, 작은 데이터 세트를 통해 파인튜닝 했을 때 성능을 크게 향상될 수 있음을 확인했다. 세 번째로, 비주류 언어 인식 능력 향상에 초점을 두어 다국어 음성인식기의 고질적 문제를 해결하고자 함에 있다.

2. 관련 연구

2.1. Transformer 를 이용한 종단형 음성인식

Transformer는 구글이 2017년에 발표한 자연어 처리 모델로 기존의 재귀신경망(recurrent neural network, RNN) 기반의 시퀀스 변환 모델의 한계를 극복하기 위해 제안되었다. Transformer는 주로 어텐션 기법에 의하여 입력 시퀀스의 각 토큰 사이의 관계를 학습하며 인코더-디코더 구조를 통해 시퀀스로 구성된 다양한 입출력에 대해서 종단형 학습 기법을 적용할 수 있다. 번역, 대화생성 등 자연어처리를 위하여 제안된 모델이지만 음성 및 영상 등 다양한 분야에 적용되어 우수한 성능을 보여주고 있다.

음성 인식 분야에서는 Transformer를 활용한 종단형 음성인식 방법은 기존의 방법들에 비해서 우수한 성능을 보이고 있으며 특히 또 다른 시퀀스 학습방법인 CTC(connectionist temporal classifcation) 기법과 함께 사용되어 학습의 수렴성을 개선할 수 있으며 이는 ESPnet 툴킷으로 공개되어 음성인식뿐만 아니라 음성합성, 통역 등 다양한 응용 분야에서 활용되고 있다(Oh et al., 2022; Watanabe et al., 2018).

2.2. Whisper 음성인식 모델

다국어 음성인식 모델은 아직까지 대규모 음성 데이터 세트 부족으로 인해 문제가 존재한다(Yadav & Sitaram, 2022). 특히 리소스가 적은 언어에 대한 성능 문제가 존재하기에 이를 해결하기 위한 여러 방법을 제안하고 검증할 필요성이 있다. 최근 OpenAI에서는 다국어 음성인식 모델인 Whisper 모델을 공개하였다. 총 약 680,000시간의 많은 양의 음성 데이터를 이용하여 학습하며, 여기에는 약 117,000시간의 비영어 다국어 음성인식 데이터와 약 125,000시간의 자동번역 데이터도 포함되어 있다.

약 117,000시간 중에 한국어 음성인식 데이터는 약 8,000시간 정도를 차지한다. Transformer의 인코더-디코더 구조를 기본으로 하여 약한 지도학습을 통해 사전 훈련한다. Whisper 모델은 기존의 음성 SSL 모델들의 한계를 극복하고, 더욱 강건한 성능을 얻기 위해 약한 지도학습(weakly supervised learning)과 멀티 태스크 학습을 통해 훈련된다. Whisper 모델은 범용 음성인식 모델로 다국어 음성인식뿐만 아니라 음성 번역과 언어 식별까지 수행할 수 있는 멀티 태스크 모델이다. 추가적으로 Whisper 모델의 특징은 음성인식이 어려운 환경에서 강하다. 악센트가 있거나 시끄러운 환경의 데이터에 대해서도 인식 및 번역 기능을 제공할 수 있는 특징을 갖고 있다.

3. 방법론

본 논문에서는 OpenAI에서 공개한 Whisper 음성인식 모델을 베이스라인으로하여, 여기에 소용량의 한국어 데이터를 이용한 파인튜닝을 통하여 한국어 음성인식 성능을 개선하고자 하였다.

3.1. 모델

공개된 Whisper 모델은 파라미터 수에 따라 표 1과 같이 여러 가지 크기의 모델이 있다. 또한 영어로만 학습된 모델과 다국어로 학습된 모델이 존재한다. 실험에 사용한 모델은 tiny, medium, base 모델로, 모두 다국어로 학습된 모델이다. 이들에 대한 인식 성능 측정을 위해 각 모델을 파인튜닝하고 추론했다. large-v2는 48G의 VRAM을 가진 GPU 카드로는 파인튜닝이 되지 않아 실험에서 제외하였다.

표 1. Whisper 모델의 크기와 scratch부터 완전히 학습된 Transformer 모델의 크기
Table 1. The size of the whisper model and the size of the Transformer model full trained from scratch

모델		파라미터 개수	모델 크기
Whisper	large-v2	1.54 B	6.17 GB
	medium	762.32 M	3.05 GB
	small	240.58 M	962.33 MB
	base	71.83 M	287.3 MB
	tiny	37.18 M	148.74 MB
Baseline Conformer		116.15 M	464.59 MB

허깅페이스(Hugging Face)에서 Whisper 모델을 ESPnet 툴킷에서 읽어올 수 있는 기능을 제공하였으며, 이를 이용하여 Whisper 모델을 ESPnet 툴킷에서 로딩하여 이 값을 초기값으로 한국어 데이터셋을 사용하여 파인튜닝하여 성능 개선 정도를 측정하였다.

또한 표 1의 Baseline Conformer는 기본 종단형 음성인식 모델을 사전학습 모델 없이 랜덤 초기값으로부터 훈련하여 이를 앞의 파인튜닝결과와 비교하기 위한 모델이다. 비교에 사용한 모델로는 ESPnet에서 제공하는 recipe 중 KsponSpeech를 훈련하는 recipe에서 제공하는 모델로 선택하였다. 훈련에 사용되는 데이터가 Whisper 모델에 비해서는 소량이므로 모델의 크기는 Whisper 모델의 large 모델 또는 medium 모델에 비하여 매우 작으며, Whisper base 모델과 small 모델의 중간정도의 크기를 갖는다. 이 모델은 12개의 Conformer 인코더 층과 8개의 Transformer 디코더 층을 가지며, 내부 파라미터로 2,048의 hidden dimensions, 256 attention dimension, 8 attention heads를 가진다. 또한 CTC 손실 함수는 0.3의 가중치로 Attention 손실 함수와 같이 결합하여 학습하였다.

3.2. 데이터셋

실험에 Whisper 모델을 파인튜닝하기 위해 KsponSpeech 데이터셋을 사용했다. KsponSpeech는 약 1,000시간의 한국어 비정형 음성 데이터 세트이며, 학습, 개발 및 평가셋으로 세트가 분할되어 있다(Bang et al., 2020).

Transformer 모델을 초기값부터 학습하여 비교하기 위해 동일한 1,000시간의 데이터셋을 사용하였다. 그리고 더 많은 데이터를 사용함에 따른 성능 변화를 확인하기 위해 추가적으로 AIHub 방송 컨텐츠 한국어 데이터를 추가로 사용했다(AiHub, 2021). AIHub 데이터는 다양한 한국 방송 컨텐츠에서 수집되었고, 수동으로 기록되었다. 데이터 길이는 약 10,000시간 이상에 달한다. 전사에 일반 구두점 기호(“?.!,) 이외에 특수 문자가 포함된 발화 데이터는 오류를 포함하거나 그 데이터를 위한 특수 처리가 필요할 수 있어서 사전에 제거했다. 예를 들어 ‘*’ 기호는 개인정보보호를 위한 마스킹에 사용된 기호이다. 너무 짧거나 긴 발언은 제외하여 총 10,000시간의 음성 데이터 중 약 6,500시간만 사용되었다.

훈련된 모델을 테스트하기 위해서 KsponSpeech 데이터셋의 평가 분할 세트인 kspon-evalother과 kspon-evalclean을 평가 데이터 세트로 사용하였다. 그리고 학습에 사용된 데이터와 성격이 다른 데이터 세트에 대한 평가를 위해 내부 한국어 평가 데이터 세트로 spon-bcast, spon-debate, spon-present를 사용했다. 각각 방송, 토론, 발표의 형식으로 자연스럽게 발성한 데이터셋이다. 각 데이터셋의 크기는 표 2에서 확인할 수 있다.

또한, Whisper 모델은 기본적으로 다국어 모델로서 영어와 한국어 모두 인식할 수 있다. 우리는 한국어 음성으로 파인튜닝 후에도 영어 인식률이 양호한지 확인하기 위해 LibriSpeech 데이터셋의 testother 데이터를 사용하여 인식률을 측정했다(Panayotov et al., 2015).

표 2. 학습된 모델 평가에 사용되는 테스트 데이터셋
Table 2. Test datasets used for evaluation of trained models

테스트 데이터셋	발화 수	시간
kspon-evalclean	3,000	3.64
kspon-evalother	3,000	3.80
spon-bcast	891	2.14
spon-debate	1,308	1.65
spon-present	1,184	1.47
libri-testother	2,939	5.34

3.3. 평가지표

연구에서는 평가 기법인 음절 오류율(character error rate, CER)을 사용하여 한국어 음성인식의 성능을 평가했다. 한글은 띄어쓰기가 모호한 부분이 많기 때문에 단어 오류율(WER)보다는 CER을 사용하는 것이 음성인식기의 성능 평가에 적합하다. 정답 전사와 각 모델의 인식 결과로부터 편집 거리를 계산하여 인식률을 측정하였고, 공백 문자는 고려하지 않았다. 모든 문장부호는 전사와 인식 결과 모두에서 제거되었다. 또한, 평가 시 간혹 영어 알파벳과 숫자가 한글 문자로 표기되거나 인식되는 등 텍스트 정규화 문제가 남아있음을 유의해야 한다. 영어 음성 데이터셋인 LibriSpeech의 테스트셋(libri-testother)의 경우 WER을 계산하여 인식률을 측정했다.

아래의 두 수식은 각각 CER과 WER을 계산하는 수식이다.

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

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

(1)에서 S_c는 인식 결과에서 잘못 대체(substitution)된 글자 수, D_c는 인식 결과에서 삭제(deletion)된 글자 수, I_c는 인식 결과에 삽입(insertion)된 글자 수, N_c는 정답 텍스트의 총 글자 수이다. 이 방식으로 글자 오류 비율을 계산할 수 있다. (2)의 경우, (1)에서 글자 수였던 기준을 단어 수로 바꾸어 계산하는 방식이다. 즉, 단어 오류 비율을 계산하는 것이다.

4. 실험

4.1. 실험환경 설정

Whisper 모델을 파인튜닝하면서 모델 크기별로 AdamW optimizer와 Warm-Up scheduler를 사용하여 학습률이 1e-05, weight decay를 0.01로 설정하였다. 그림 1과 같이 수렴하기에 충분한 조건으로 최대 epoch 수는 3으로 설정하였다. 8개의 NVIDIA A6000 GPU가 장착된 기계를 사용하여 모델 크기별로 각각 0.48시간, 0.62시간, 26.07시간이 소요되었다.

사전훈련된 모델 없이 훈련하는 Transformer 모델의 경우 총 세 가지 종류의 학습 데이터 구성을 이용하여 학습되었다. 1) Whisper 모델의 파인튜닝 때 사용된 것과 동일한 KsponSpeech 데이터 1,000시간 데이터로 훈련한다. 2) 데이터의 증가에 따른 성능의 변화를 관찰하기 위하여 6,500 시간의 AIHub 데이터를 사용하여 훈련하였고, 3) 훈련데이터와 평가데이터의 도메인 유사도를 반영하기 위하여 두 학습 데이터 세트를 모두 사용한 7,500시간의 학습 데이터를 사용하여 훈련한다. 각 3개 모델에 대해 학습률 0.002, weight-decay 0.01의 Adam optimizer 및 Warm-Up scheduler를 사용했다. 모델은 early stop 없이 최대 35 epoch까지 학습되며, 마지막 5개의 모델들에 대해 평균을 냈다. 8개의 NVIDIA A6000 GPU로 3개의 모델을 각각 훈련하는 데에 걸린 시간은 8.5시간, 37.0시간, 39.2시간이 소요되었다.

4.2. Whisper 모델의 Zero-Shot 평가

먼저, Whisper 모델의 zero-shot 평가를 진행했다. 앞서 설명한 바와 같이 Whisper 모델은 대용량의 데이터를 사용하여 파인튜닝없이 사용하도록 공개된 모델이다. 따라서 zero-shot 평가도 우수한 성능이 나올 것으로 예상할 수 있다. 단 Whisper의 모델을 ESPnet의 모델로 변환하는 과정에서 기존 모델의 성능이 달라 질 수 있다. 파인튜닝을 적용하지 않은 모델을 0 epoch 모델이라고 칭하겠다. 0 epoch 모델을 ESPnet을 사용하여 inference 한 결과, CER이 모든 평가셋에 대하여 60% 이상으로 높은 오류율을 보여주었다(그림 1의 0 epoch 참고). 이는 음성의 끝을 나타내는 <eos> 토큰이 제대로 인식되지 않아, 마지막 부분의 전사가 반복적으로 추가되는 삽입 오류가 발생했기 때문으로 추정된다.

Whisper 모델은 모델과 같이 제공하는 인식기를 사용하여 command-line으로 음성 파일을 텍스트로 전사하는 기능을 지원한다. 이를 stand-alone 방식이라고 칭하겠다.

0 epoch 모델의 인식 오류율이 높아 stand-alone 방식으로도 inference를 진행했다. 이 방법에서는 insertion 오류가 훨씬 작게 측정되었다. Stand-alone 방식의 방법은 참고 문헌 OpenAi github에서 볼 수 있다(OpenAi, 2023).

표 3. ESPnet 기반 Whisper 모델로 inference한 결과와 Whisper 모델을 command-line으로 inference한 결과

Table 3. Results of inferring with the ESPnet-based Whisper model and command-line of the Whisper model

ESPnet 기반 Whisper 모델 (0 epoch)	A_006_0003 (ko) 그리고 특히 무석 이월 정리금 문제 뿐만 아니라 대학 등록금이 무섭가 상승률을 할 수 있는 상향다는 것이 통계 자료에서 나타나고 있고 있고 그리고 특히 무섭 이월 정리금 문제 뿐만 아니라 대학 등록금이 무섭가 상승률을 할 수 있는 상향다는 것이 통계 자료에서 나타나고 있고 A_006_0004 (ko) 이제 주로 2배에 3배 정도 훨씬 더 상의하는 수준으로 등록금 인상이 진행되고 있습니다. 있습니다. 1688-142285-0000 (en) "There's eye and they say, in all our blood, and a grain or tube, perhaps, is good. But his, he makes me harshly feel, has got a little too much of steel." An on. An on
Whisper (stand-alone)	A_006_0003 (ko) 그리고 특히 무섭 2월 정리금 문제 뿐만 아니라 대학 등록금이 물가 상승률을 훨씬 상의한다는 것이 좀 통계자료에서 나타나고 있고 A_006_0004 (ko) 주로 2배 3배 정도 훨씬 더 상의하는 수준으로 등록금 인상이 진행되고 있습니다. 1688-142285-0000 (en) There's I and they say, in all our blood, and a grain or tube, perhaps, is good. But here's, he makes me harshly feel, has got a little too much of steel, and on.

표 3은 0 epoch 모델을 ESPnet으로 인식한 결과와 stand-alone 방식의 Whisper 모델 각각의 인식 결과를 예시를 보여준다. 모두 beam size는 3으로 설정하였다. 결과에서 보듯이 ESPnet으로 인식하는 경우 문장의 끝을 정확히 찾지 못하고 잘못된 단어가 삽입됨을 확인할 수 있었다.

표 4는 stand-alone 방식의 Whisper 모델 크기별 CER을 보여준다. Whisper 모델의 크기가 클수록 인식 성능이 우수함을 알

수 있으며 medium 모델과 base 모델은 큰 성능 차이를 보이나, large-v2 모델과 medium 모델의 성능차이는 크지 않음을 확인할 수 있다.

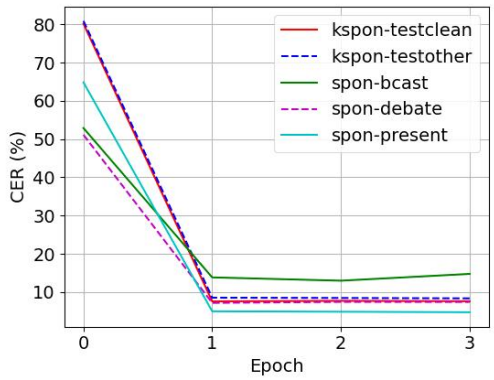
표 4. Command-line으로 inference한 Whisper 모델의 크기별 음절 오류율(CER, %)[libri-other의 경우, 단어 오류율(WER, %)로 측정됨]
Table 4. CER for each size of the Whisper model inferred by command-line (for libri-other, measured by WER)

테스트 데이터셋	large-v2	medium	base	tiny
kspon-evalclean	10.83	11.94	22.85	33.57
kspon-evalother	10.83	11.96	21.12	30.90
spon-bcast	13.14	14.03	21.30	28.49
spon-debate	10.25	11.48	18.65	25.97
spon-present	11.14	11.49	16.76	21.91
libri-testother	6.31	6.74	11.95	15.79

CER, character error rate; WER, word error rate.

4.3. Whisper 모델의 파인튜닝

본 연구 KsponSpeech를 이용하여 Whisper 모델의 medium, base 및 tiny 모델을 파인튜닝했다. 그림 1은 3 epoch까지 파인튜닝한 Whisper medium 모델의 각 epoch별 CER을 보여준다. Whisper 모델은 1 epoch만 파인튜닝하더라도 zero-shot 평가보다 좋은 인식 성능을 보여주었다. 0 epoch의 경우 앞서 언급한 삽입 오류 문제 때문에 CER이 높게 측정되었다. 최종 모델은 1, 2, 3 epoch의 모델을 평균하여 사용하였다.



CER, character error rate.

그림 1. KsponSpeech로 파인튜닝한 Whisper medium 모델의 각 epoch별 CER
Figure 1. CER for each epoch of the Whisper medium model fine-tuned with KsponSpeech

표 5는 각각 Whisper tiny, base, medium 모델로부터 파인튜닝한 한국어 음성 인식 성능에 대한 CER을 보여준다. 1,000시간이라는 비교적 적은 훈련데이터를 사용하여 파인튜닝했음에도 성능이 크게 개선됨을 확인하였다. 특히 데이터의 성격이 파인튜닝에 사용한 데이터와 유사한 kspon-evalclean 및 kspon-evalother 평가데이터뿐만 아니라 다른 평가셋에 대해서도 오류감소율이 12.82%에서 최대 58.75%로 성능이 크게 개선됨을 확인할 수 있다. 다만 spon-bcast에 대해서는 zero-shot 추론에 비해서 성능이 감소하였다. 또한 파인튜닝한 Whisper 모델 역시 크기가 클수록 인식 성능이 우수함을 알 수 있다.

표 5. Whisper medium, base 및 tiny모델을 파인튜닝한 이후 한국어 평가셋에 대한 음절 오류율(CER, %) 및 zero-shot 추론 대비 오류 감소율(%)
Table 5. CER (%) for each size of the Whisper model after fine-tuning and error reduction ratio (ERR) compared to zero-shot inference (%)

테스트 데이터셋	CER (%)			ERR (%)		
	medi	base	tiny	medi	base	tiny
kspon-evalclean	7.61	12.99	15.94	36.26	43.15	52.52
kspon-evalother	8.36	13.68	17.67	30.10	35.23	42.82
spon-bcast	14.75	20.91	29.44	−5.13	1.83	−3.33
spon-debate	7.42	16.01	22.64	35.37	14.16	12.82
spon-present	4.74	9.35	12.87	58.75	44.21	41.26

CER, character error rate.

4.4. 파인튜닝 모델의 다국어 인식 성능

Whisper 모델은 96개 언어에 대한 인식 및 번역이 가능한 다국어 모델이며, 언어의 종류는 학습 및 추론 과정에서 프롬프트를 제공함으로써 지정할 수 있다. 본 실험에서는 한국어의 파인튜닝을 위해서 프롬프트를 한국어로 지정하였으며, 이에 따라 한국어 인식결과가 개선됨을 확인하였다. 한국어 데이터로 파인튜닝된 모델의 한국어 이외의 언어에 대한 인식 성능을 확인하기 위하여 영어 데이터로 평가를 추가로 진행하였다. 표 6은 Whisper의 zero-shot 인식결과와 파인튜닝된 모델에 대한 인식결과를 보여준다. 인식결과에서 보듯이 파인튜닝된 모델의 결과는 매우 열화되었으며 특히 100% 이상의 WER은 삽입 오류가 많이 발생하였음을 의미한다. 또한 프롬프트로 영어 인식을 지정하였음에도 한국어로 인식되는 경우도 다수 발생하였다. 이는 프롬프트로 한국어를 지정하여 파인튜닝을 하였으나 다른 언어의 인식성능도 큰 영향을 받음을 의미한다.

표 6. 한국어로 파인튜닝된 모델에 대한 영어 인식 성능 단어 오류율(WER, %)
Table 6. WER for zero-shot inference and fined-tuned Whisper model of each size

모델	medium	base	tiny
Zero-shot 추론	6.74	11.95	15.79
한국어 파인튜닝 모델	57.56	42.05	113.17

WER, word error rate.

4.5. 파인튜닝된 Whisper 모델과 전체 학습된 모델 간의 성능 비교

사전학습모델이 음성인식 성능에 미치는 영향을 확인하기 위하여 파인튜닝된 모델의 성능과 사전학습 모델 없이 임의 초기값으로부터 완전 학습된 모델의 성능을 비교해보았다. 완전 학습모델은 3.1절에서 설명한 Baseline Transformer 모델 구조를

사용하였으며, ESPnet을 사용하여 학습하였다. 학습된 모델을 이용한 평가 결과는 표 7에 제시하였다. 세 가지 종류의 훈련데이터를 사용하여 학습을 진행하였다. 먼저 1,000시간 전체학습은 파인튜닝에 사용한 것과 완전히 동일한 데이터셋으로 전체학습하였다. 두 번째로 6,500시간 전체학습은 훈련데이터의 증가에 따른 전체학습의 영향을 파악하기 위하여 AIHub 데이터를 사용하여 학습한 모델이다. 이 모델의 경우 kspon-evalclean 평가셋에 대해서는 성능이 감소함을 확인하였다. 이는 훈련데이터와 평가데이터의 유사성에서 기인한 것으로 두 데이터셋을 합친 7,500시간 전체 학습 모델의 경우 이러한 문제가 해소되었다. 7,500시간으로 전체학습한 모델의 경우 모든 평가셋에 대하여 1,000시간 학습 모델보다 우수한 성능을 보인다.

1,000시간 데이터만을 사용하여 파인튜닝된 medium 모델과 비교해보면, 1,000시간 및 6,500시간의 데이터를 이용하여 훈련한 모델과 비교하여서는 대부분의 경우 파인튜닝된 Whisper 모델이 우수한 성능을 보였다. 다만 7,500시간의 데이터를 사용하는 경우에는 전체학습 모델이 파인튜닝된 Whisper medium 모델보다 대부분의 경우 우수한 성능을 보였다. 학습에 소요된 시간은 파인튜닝의 경우 약 26시간, 전체학습의경우 약 39.2시간으로 학습 비용도 파인튜닝의 경우가 약 2/3 정도 적게 소요되었다. 이로서 사전학습된 모델을 사용하여 소용량의 데이터로 파인튜닝하는 방법의 효용성을 확인할 수 있다.

표 7. 1,000시간, 6,500시간, 7,500시간 데이터를 사용하여 전체 학습된 Transformer 모델과 파인튜닝된 Whisper medium 모델의 성능 비교(음절 오류율, %)
Table 7. CER for full trained baseline transformer model and comparison to fine-tuned Whiser medium model (CER, %)

테스트 데이터셋	1,000시간 전체학습	6,500시간 전체학습	7,500시간 전체학습	파인튜닝 medium 모델
kspon-evalclean	8.80	9.84	7.60	7.61
kspon-evalother	9.74	9.44	8.29	8.36
spon-bcast	27.04	11.85	11.53	14.75
spon-debate	13.43	7.12	6.85	7.42
spon-present	6.49	7.81	6.98	4.74

CER, character error rate.

5. 결론

본 연구는 한국어 데이터 세트에서 파인튜닝된 Whisper 음성인식 모델의 성능을 평가한다. 이와 파인튜닝되지 않은 모델 성능 및 사전학습 모델 없이 전체 훈련된 Transformer 모델 성능간 비교를 진행했다. 결과적으로 소소의 학습 epoch만으로 Whisper 모델의 파인튜닝 작업이 기존 모델 및 전체학습된 모델에 비교하여 현저하게 우수한 성능을 보여준다는 것을 확인할 수 있다. 이는 공개된 대규모 음성인식 모델을 활용하여 도메인에 맞는 소수 데이터로 파인튜닝함으로써 다양한 응용 분야에서 고성능의 음성인식기를 활용할 수 있는 가능성을 시사한다.

하지만 한국어 데이터셋으로 파인튜닝된 Whisper 모델을 영어 데이터셋인 LibriSpeech에서 테스트했을 때 WER이 매우 열화됨을 확인할 수 있다. 이를 통해 한국어에 파인튜닝된 모델이 더 이상 영어 인식기나 다국어 인식기로 사용되기 어려움을 보여준다. 또한 사전학습된 모델의 사이즈가 방대하여 실시간 추론이 어려운 문제도 있다. 향후 연구에서는 이 문제를 극복하여 다국어 인식에 대한 파인튜닝 및 실시간 추론 최적화를 통한 연구에 초점을 맞추고자 한다.

References

AiHub (2021). Aihub broadcast content korean speech recognition data. Retrieved from https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=463

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020, December). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proceedings of the Advances in Neural Information Processing Systems* (pp. 12449-12460). Online Conference.

Bang, J. U., Yun, S., Kim, S. H., Choi, M. Y., Lee, M. K., Kim, Y. J., Kim, D. H., ... Kim, S. H. (2020). KsponSpeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences*, *10*(19), 6936.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016, March). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4960-4964). Shanghai, China.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, July). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning* (pp. 1597-1607). Online Conference.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Retrieved from https://arxiv.org/abs/1810.04805

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning* (pp. 369-376). Pittsburgh, PA.

Graves, A., & Jaitly, N. (2014, June). Towards end-to-end speech recognition with recurrent neural networks. *Proceedings of the 31st International Conference on Machine Learning* (pp. 1764-1772). Beijing, China.

Hadsell, R., Chopra, S., & LeCun, Y. (2006, June). Dimensionality reduction by learning an invariant mapping. *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (pp. 1735-1742). New York, NY.

Nam, K. (2019). A study on processing of speech recognition Korean words. *The Journal of the Convergence on Culture Technology,*

5(4), 407-412.

Oh, Y. R., Park, K., & Park, J. G. (2022). Fast offline transformer-based end-to-end automatic speech recognition for real-world applications. *ETRI Journal*, *44*(3), 476-490.

OpenAi (2023). Openai/whisper. Retrieved from https://github.com/openai/whisper

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). Librispeech: An ASR corpus based on public domain audio books. *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206-5210). South Brisbane, Australia.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning* (pp. 28492-28518). Honolulu, HI.

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019, September). wav2vec: Unsupervised pre-training for speech recognition. *Proceedings of the Interspeech 2019* (pp. 3465-3469). Graz, Austria.

Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019, July). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* (pp. 6558-6569). Florence, Italy.

van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. Retrieved from https://doi.org/10.48550/arxiv.1807.03748

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., ... Polosukhin, I. (2017, December). Attention is all you need. *Proceedings of the Advances in Neural Information Processing Systems.* Long Beach, CA.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., ... Ochiai, T. (2018). ESPnet: End-to-end speech processing toolkit. Retrieved from https://doi.org/10.48550/arXiv.1804.00015

Yadav, H., & Sitaram, S. (2022, June). A survey of multilingual models for automatic speech recognition. *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 5071-5079). Marseille, France.

• **오창한 (Changhan Oh)**
한국전자통신연구원 복합지능연구실 UST 학생연구원
대전광역시 유성구 가정로 218
Tel: 042-860-6114
Email:ochh508@etri.re.kr
관심분야: 인공지능, 음성인식, 복합지능

• **김청빈 (Cheongbin Kim)**
한국전자통신연구원 복합지능연구실 학연과정
대전광역시 유성구 가정로 218
Tel: 042-860-6114
Email: chung1530@etri.re.kr
관심분야: 인공지능, 음성인식, 복합지능

• **박기영 (Kiyoung Park)** 교신저자
한국전자통신연구원 복합지능연구실 책임연구원
대전광역시 유성구 가정로 218
Tel: 042-860-1228
Email: pkyoung@etri.re.kr
관심분야: 인공지능, 음성인식, 복합지능

대형 사전훈련 모델의 파인튜닝을 통한 강건한 한국어 음성인식 모델 구축*

오 창 한 · 김 청 빈 · 박 기 영

한국전자통신연구원 복합지능연구실

국문초록

자동 음성 인식(automatic speech recognition, ASR)은 딥러닝 기반 접근 방식으로 혁신되었으며, 그중에서도 자기 지도 학습 방법이 특히 효과적일 수 있음이 입증되고 있다. 본 연구에서는 다국어 ASR 시스템인 OpenAI의 Whisper 모델의 한국어 성능을 향상시키는 것을 목표하여 다국어 음성인식 시스템에서의 비주류 언어의 성능 문제를 개선하고자 한다. Whisper는 대용량 웹 음성 데이터 코퍼스(약 68만 시간)에서 사전 학습되었으며 주요 언어에 대한 강력한 인식 성능을 입증했다. 그러나 훈련 중 주요 언어가 아닌 한국어와 같은 언어를 인식하는 데 어려움을 겪을 수 있다. 우리는 약 1,000시간의 한국어 음성으로 구성된 추가 데이터 세트로 Whisper 모델을 파인튜닝하여 이 문제를 해결한다. 또한 동일한 데이터 세트를 사용하여 전체 훈련된 Transformer 모델을 베이스 라인으로 선정하여 성능을 비교한다. 실험 결과를 통해 Whisper 모델을 파인튜닝하면 문자 오류율(character error rate, CER) 측면에서 한국어 음성 인식 기능이 크게 향상되었음을 확인할 수 있다. 특히 모델 크기가 증가함에 따라 성능이 향상되는 경향을 포착하였다. 그러나 Whisper 모델의 영어 성능은 파인튜닝 후 성능이 저하됨을 확인하여 강력한 다국어 모델을 개발하기 위한 추가 연구의 필요성을 확인할 수 있었다. 추가적으로 우리의 연구는 한국어 음성인식 애플리케이션에 파인튜닝된 Whisper 모델을 활용할 수 있는 가능성을 확인할 수 있다. 향후 연구는 실시간 추론을 위한 다국어 인식과 최적화에 초점을 맞춰 실용적 연구를 이어갈 수 있겠다.

핵심어: 딥러닝, 머신러닝, 음성인식, 음성공학

참고문헌

남기훈(2019). 한글 단어의 음성 인식 처리에 관한 연구. *문화기술융합*, *5*(4), 407-412.

* 본 연구는 한국 정부(MSIT)가 지원하는 정보통신기술기획평가원(IITP)의 보조금 지원을 받아 수행된 연구입니다(과제번호: 2022-0-00989, 다화자 대화 모델링을 위한 인공지능 기술 개발).