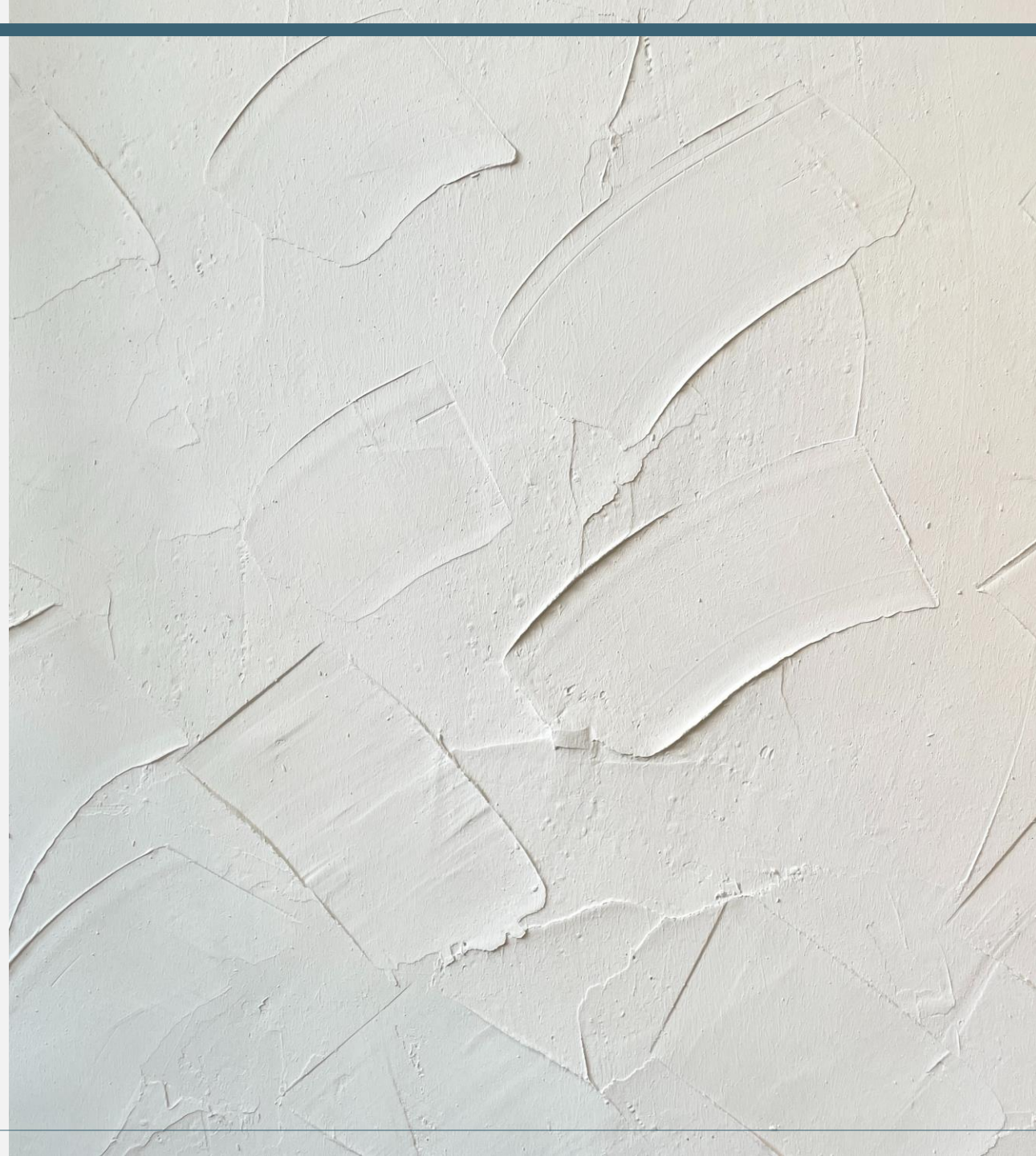
Building robust Korean speech recognition model By fine-tuning large pretrained model

Changhan Oh. Cheongbin Kim. Kiyoung Park, ETRI, 한국음성학회지, 2023

목차

table of contents



서론

- 자동음성인식(ASR) 기술은 딥러닝 기반 종단형 모델 접근법으로 혁신되고 있음 · 과거 통계 기반 시스템 대체 · 인공신경망 기반 음성-언어 패턴 분류 학습

- 최근 Transformer 모델 기반 대규모 사전학습 모델을 파인튜닝하는 방식 확산 · OpenAI Whisper 모델: 다국어 인식 가능하나 한국어 등 비주류 언어는 오류율 높음

- 연구 목표: Whisper 모델의 한국어 성능 개선을 위해 한국어 데이터로 파인튜닝

관련연구

1) Transformer 기반 종단형 음성인식

- Transformer: 2017년 구글에서 제안한 자연어처리 모델 · 기존 RNN 기반 모델의 한계 극복 · 어텐션 기법으로 입력 시퀀스 토큰 간 관계 학습

- 음성인식에서 Transformer + CTC 결합 활용 · 수렴성 개선 및 다양한 응용 분야 활용

2) Whisper 음성인식 모델

- 다국어 음성인식 모델의 데이터 문제 존재 · 리소스 적은 언어에 대한 성능 이슈

- Whisper: OpenAI 다국어 음성인식 모델 · 68만 시간 웹 음성데이터로 약한 지도학습 · 비영어 11.7만 시간 데이터 포함(한국어 0.8만 시간) · 다국어 인식, 번역, 언어 식별 등 멀티태스크 수행 · 악센트, 시끄러운 환경에서도 강건한 성능

방법론

- 베이스라인: OpenAI Whisper모델
- 약 1,000시간 한국어 데이터로 Whisper 파인 튜닝

Whisper 모델의 크기와 scratch부터 완전히 학습된 Transformer 모델의 크기

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

- Baseline Conformer는 기본 종단형 음성인식 모델 을 사전학습 모델 없이 랜덤 초기값으로부터 훈련하여 이를 앞 의 파인튜닝결과와 비교하기 위한 모델

데이터셋

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

- 파인튜닝: **KsponSpeech** 약 **1,000**시간 데이터
- 평가: **KsponSpeech** 평가셋 외 · **spon-bcast, spon-debate, spon-present** (방송/토론/발표 데이터)
- **Transformer** 전체학습에 **AIHub 6,500**시간 추가 데이터 활용

평가지표

- 한국어 평가: 문자오류율(CER) 사용 · 편집거리 기반 정답 전사와 인식결과 간 오류율 측정
- 영어 평가(LibriSpeech): 단어오류율(WER) 사용

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

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

실험

1) 환경설정

•Whisper 모델 파인튜닝 · AdamW 옵티마이저, Warm-Up 스케줄러 사용 · 학습률 1e-05, weight decay 0.01 · 최대 3 epoch 학습 · 8개 NVIDIA A6000 GPU 사용

•Transformer 전체 학습 · Adam 옵티마이저, Warm-Up 스케줄러 · 학습률 0.002, weight decay 0.01· 최대 35 epoch, 마지막 5개 모델 평균 · 8개 NVIDIA A6000 GPU 사용

2) 결과 요약

• Whisper Zero-Shot 평가 · Command-line 기반 추론 시 11-14% CER 수준 (Medium 모델 기준)

•Whisper 모델 파인튜닝 · 1,000시간 한국어 데이터로 파인튜닝 · 모든 모델 크기에서 큰 폭 성능 개선 - Medium 모델 기준 최대 58.75% 오류 감소 · 모델 크기가 클수록 인식 성능 우수

•다국어 인식 성능 · 한국어 데이터로 파인튜닝 후 영어 WER 크게 열화 - Medium 모델 기준 6.74% → 57.56%

•전체 학습 모델 비교 · 대체로 파인튜닝 모델이 전체 학습 모델보다 우수 - 7,500시간 학습 모델의 경우 파인튜닝보다 약간 좋음 · 파인튜닝 학습 비용이 전체 학습의 2/3 수준