Whisper Model

Robust Speech Recognition via Large-Scale Weak Supervision

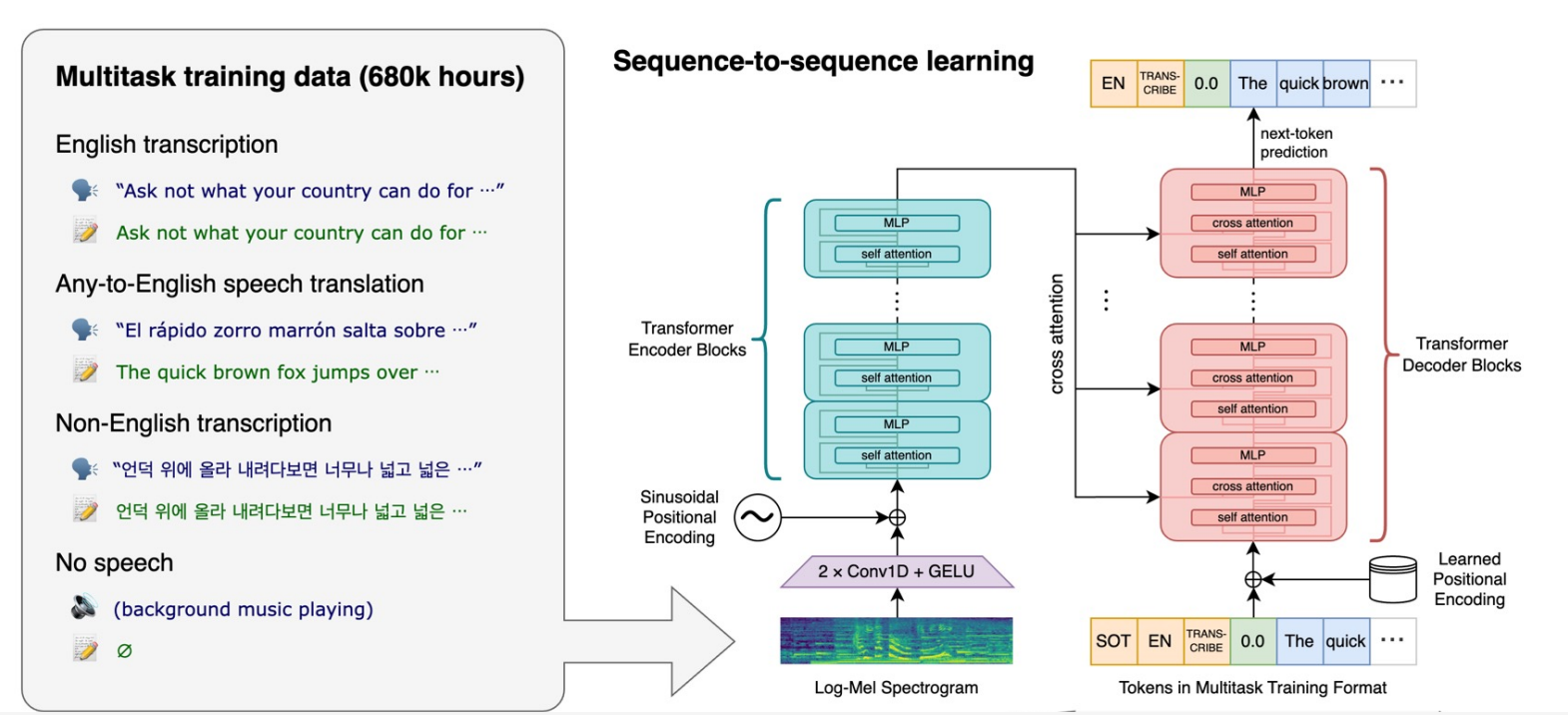
발표자: 이세영

2023. 10. 06

목차

1. Introduction

Whisper 모델?



Github: https://github.com/openai/whisper

- 2022. 09 공개된 모델
- 음성인식 end-to-end 모델
- 다양한 언어 가능 (en, ko, ja, zh 등)
- 다양한 Task 가능
  음성 전사, 번역, 언어 감지 등
- 특히, 한국어 가능 모델 중 뛰어난 성능을 보임.

1. Introduction

- Return Zero의 한국어 음성인식 리더보드

API \ 데이터셋	Avg. CER(%)	주요 영역별 회의	회의	상담	저음질 전화망	한국어 강의	KsponSpeech eval clean	KsponSpeech eval other
OpenAI Whisper	11.39	10.49	10.16	7.51	17.27	10.89	12.06	11.34
Google api v2	11.50	N/A[1]	11.62	8.37	14.11	11.48	11.82	11.59
ETRI	10.19	9.95	10.56	8.36	15.46	9.89	9.99	7.15
Naver ClovaSpeech	9.52	7.88	8.53	5.89	9.09	13.71	10.66	10.86
리턴제로	**6.18**	**6.78**	**7.27**	**3.56**	**4.66**	**7.76**	**6.61**	**6.64**
리턴제로 Whisper[2]	7.79	6.43	8.85	5.44	5.52	8.68	9.74	9.86

Github: https://github.com/rtzr/Awesome-Korean-Speech-Recognition

- API 이용 가능 모델 평가
- Whisper만 공개 모델
- ETRI 모델의 경우, 상업이용시 라이센스 필요

1. Introduction

- 다양한 사이즈의 모델 공개

Size	Parameters	English-only model	Multilingual model	Required VRAM	Relative speed
tiny	39 M	`tiny.en`	`tiny`	~1 GB	~32x
base	74 M	`base.en`	`base`	~1 GB	~16x
small	244 M	`small.en`	`small`	~2 GB	~6x
medium	769 M	`medium.en`	`medium`	~5 GB	~2x
large	1550 M	N/A	`large`	~10 GB	1x

Github: https://github.com/openai/whisper

2. Whisper

Whisper: Web-scale Supervised Pretraining for Speech Recognition(WSPSR)에서 따온 방법론 이름

Main Contribution

- simple scaling of weakly supervised pre-training
  - 다국어 데이터셋 117,113 hours, Translation 125,739 hours, 영어 음성인식 438,218 hours
- 68만 시간 labeled audio data 학습
- transcript text standardization 제안

2. Whisper

English Text

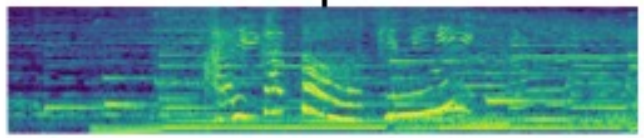
1. Remove any phrases
2. Remove any phrases
3. Remove any of the fo
4. Remove whitespace
5. Convert standard or i
6. Remove commas (,)
7. Remove periods (.)
8. Remove symbols as
   starting with M, S, or P, except period, percent, and currency symbols that may be detected in the next step.
9. Detect any numeric expressions of numbers and currencies and replace with a form using Arabic numbers, e.g. "Ten thousand dollars" → "$10000".
10. Convert British spellings into American spellings.
11. Remove remaining symbols that are not part of any numeric expressions.
12. Replace any successive whitespace characters with a space.

Non-English Text

1. Remove any phrases between matching brackets ([,]).
2. Remove any phrases between matching parentheses ((,)).
3. Replace any markers, symbols, and punctuation characters with a space, i.e. when the Unicode category of each character in the NFKC-normalized string starts with M, S, or P.
4. make the text lowercase.
5. replace any successive whitespace characters with a space.

2. Whisper

Input data: Log-Mel Spectrogram

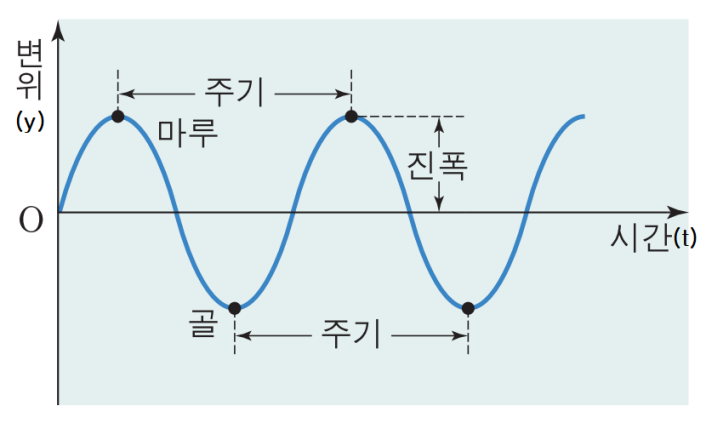


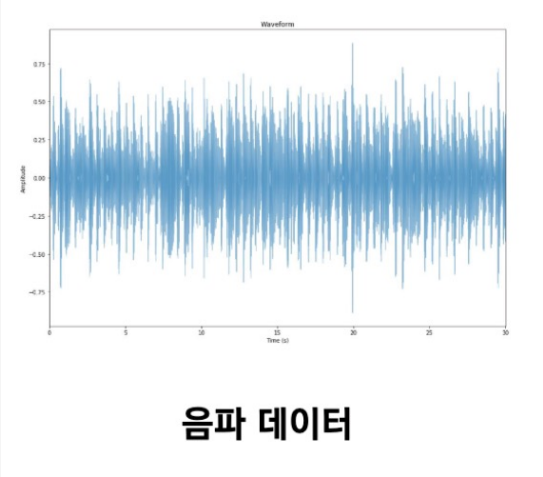
Log-Mel Spectrogram

- 음파 데이터를 사람 귀가 듣는 정보와 비슷하게 scaling 하는 과정을 거침 -> Mel spectrogram
- 연산량이 많아 normalization 기법으로 log를 취함 -> Log-Mel Spectrogram
  - Mel spectrogram을 압축하여 표현한 MFCC 형식을 사용하기도 함.
- Whisper 모델은 Log-Mel Spectrogram을 input로 사용함.

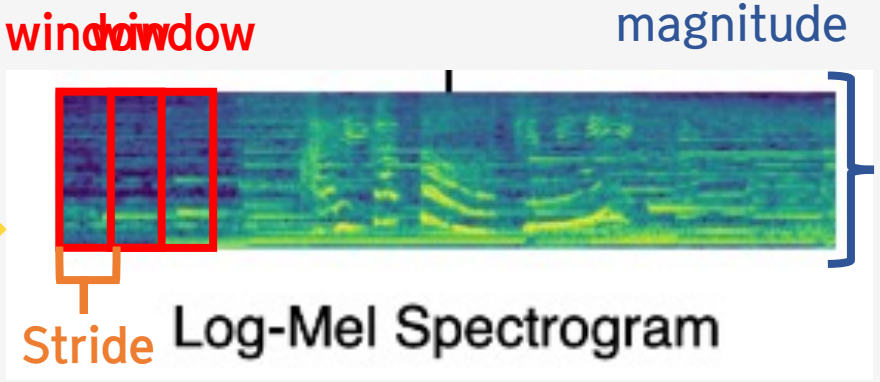
2. Whisper

Input data: Log-Mel Spectrogram





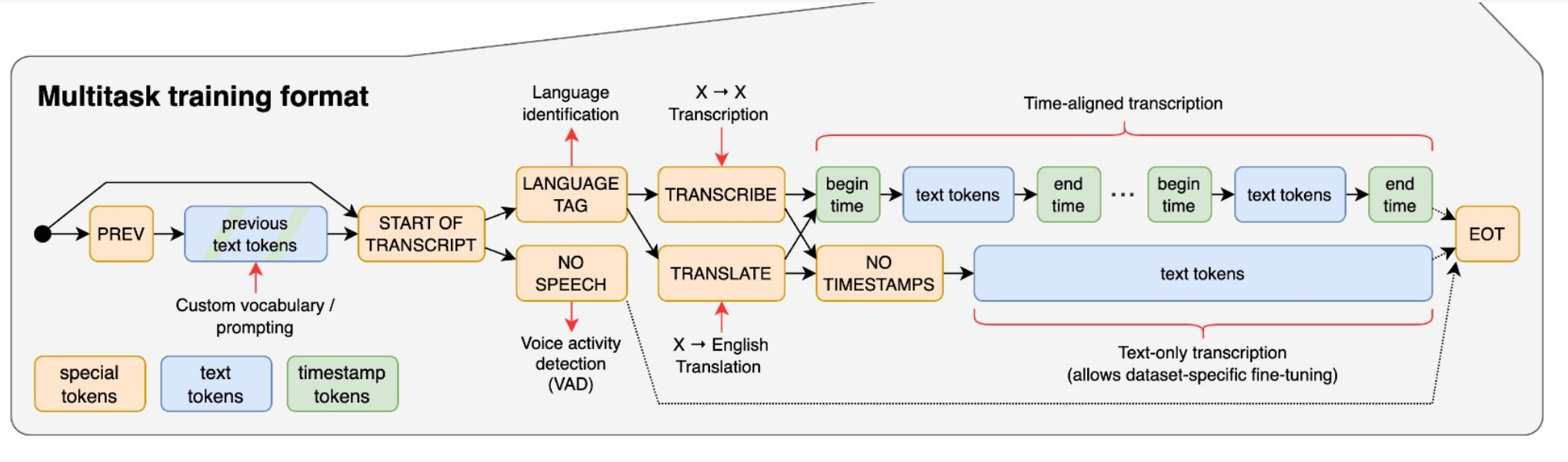
음파 데이터



- sampling rate: 16000(1초에 얼마나 많은 점을 찍을것인가)
- log-magnitude: 80-channel
- windows:25ms
- Stride:10ms

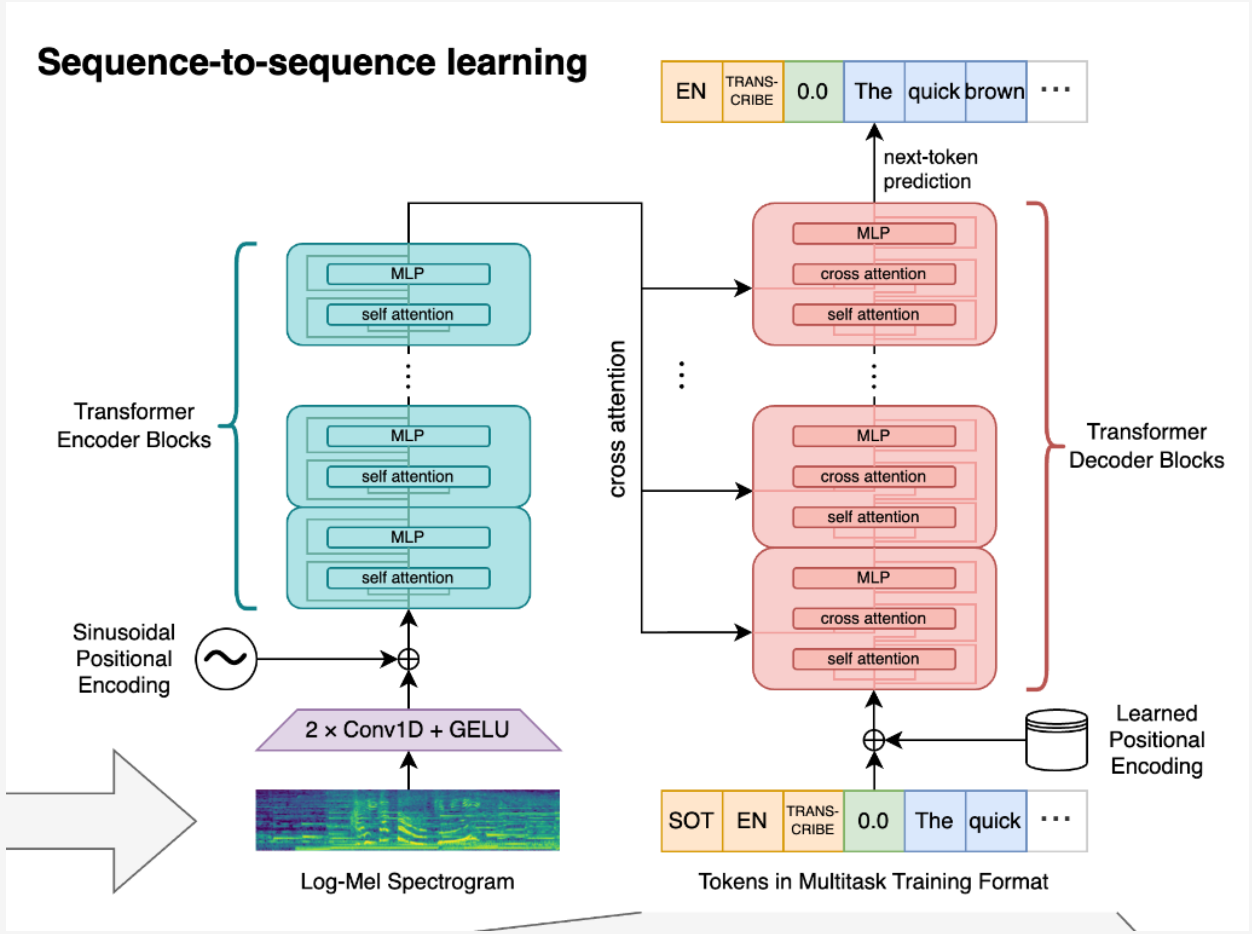
2. Whisper

Output data: Multitask training format



2. Whisper

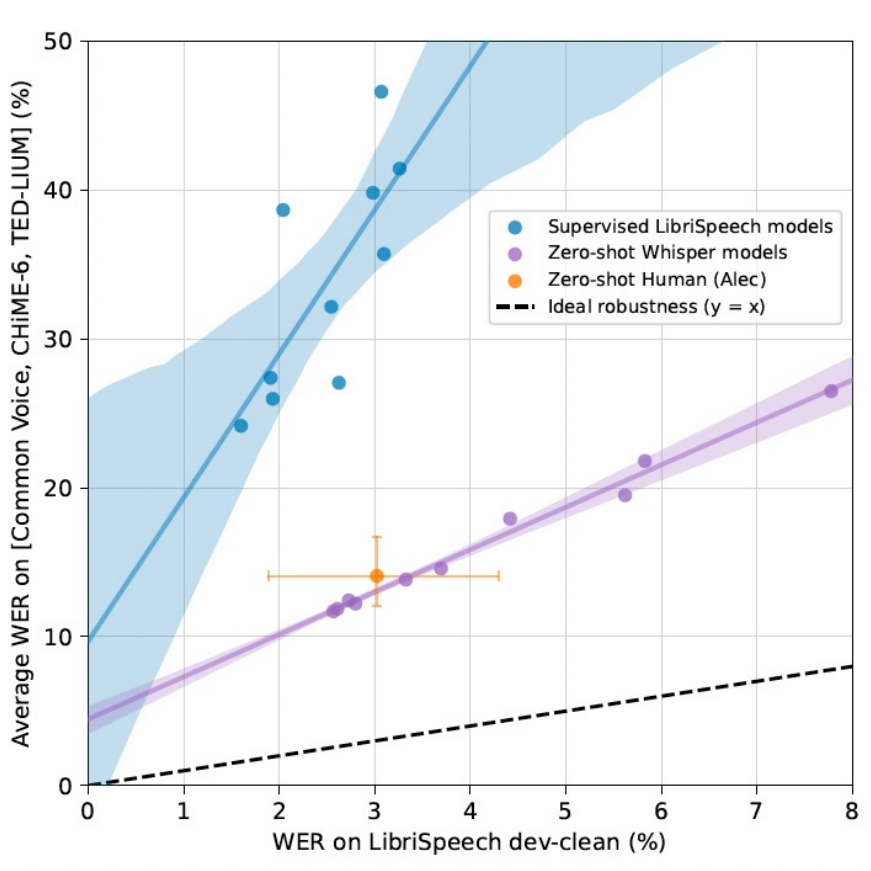
Model Architecture



- Transformers Architecture 사용
- Tokenizer: GPT-2 BPE tokenizer 사용
  - Vocab을 refit하여 사용함.

3. Experiment

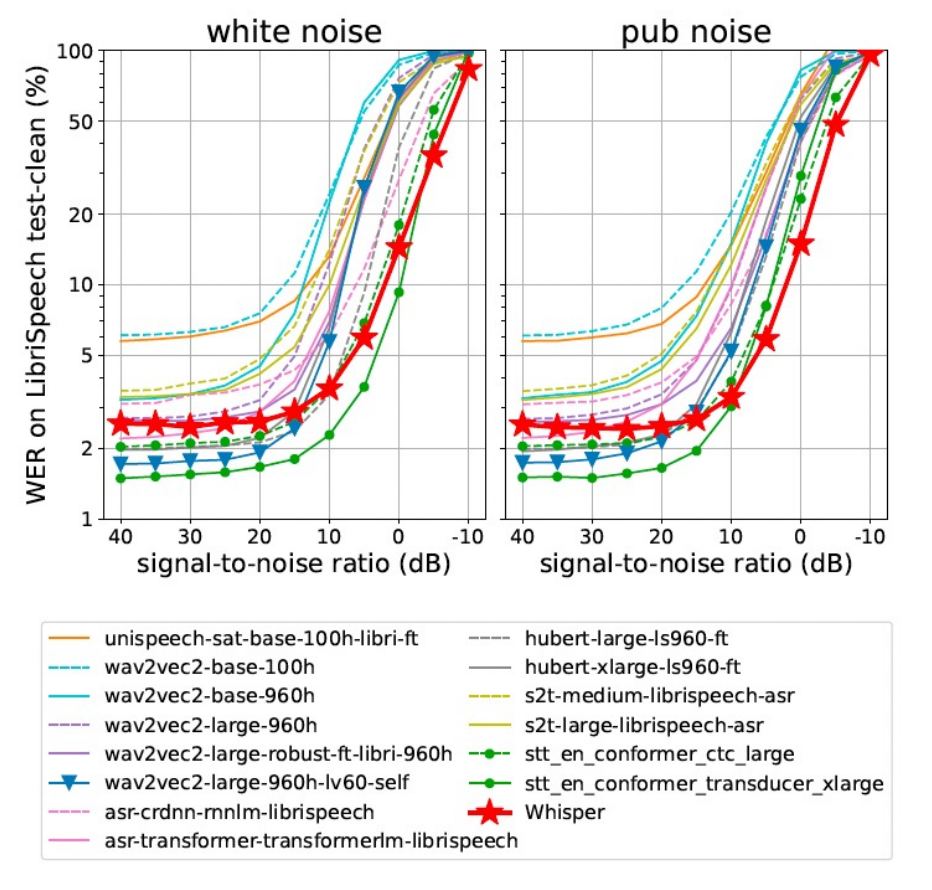
(1) Zero-shot performance



- 2019 SOTA 모델을 사용하여 LibriSpeech를 학습한 모델과 LibriSpeech를 학습하지 않은 Whisper 성능 비교
- Supervised 모델이 인간의 2배 이상의 WER을 보임.
- Zero-shot Whisper 가 인간과 유사한 성능을 보임.

3. Experiment

(2) Robustness to Additive Noise



- White noise, pub noise 를 추가한 LibriSpeech test-clean 데이터셋으로 평가함.
- 낮은 noise 환경에서는 NVIDIA의 STT models가 성능이 좋지만, noise가 강해질수록 Whisper 모델의 성능이 가장 좋다.

3. Experiment

(3) Translation

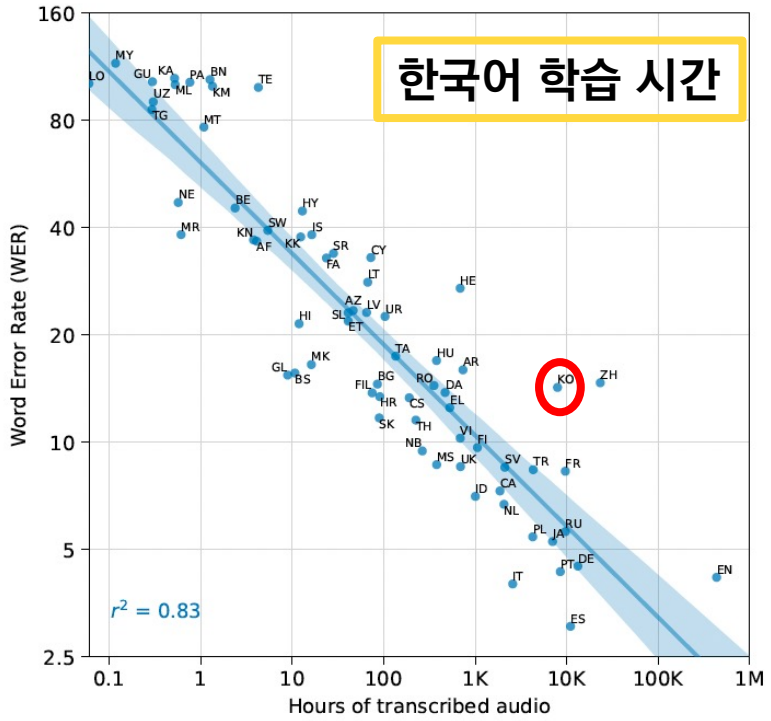


Figure 3. **Correlation of pre-training supervision amount with downstream speech recognition performance.** The amount of pre-training speech recognition data for a given language is very predictive of zero-shot performance on that language in Fleurs.

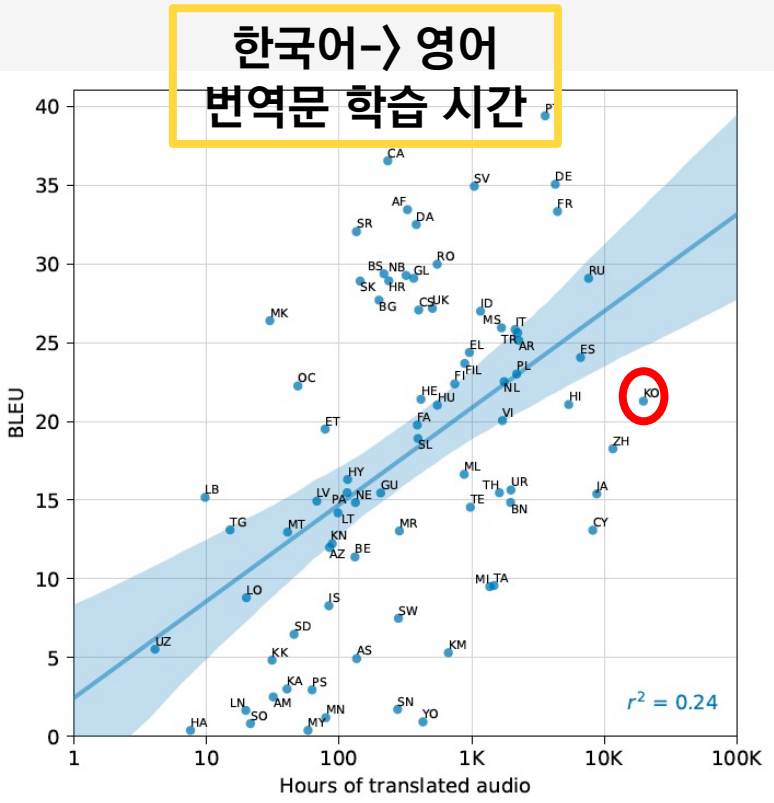


Figure 4. **Correlation of pre-training supervision amount with downstream translation performance.** The amount of pre-training translation data for a given language is only moderately predictive of Whisper's zero-shot performance on that language in Fleurs.

- 타 모델과 번역 성능 비교(Bleu)

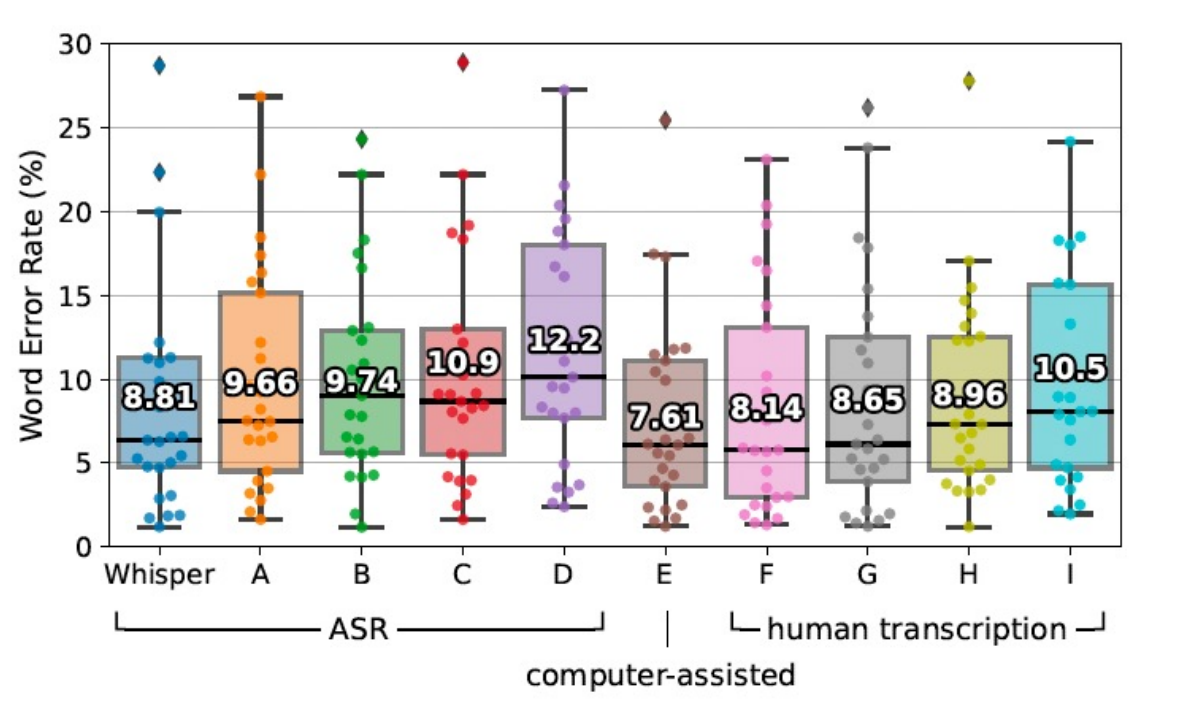
X → English	High	Mid	Low	All
XMEF-X	34.2	20.2	5.9	14.7
XLS-R (2B)	36.1	27.7	15.1	22.1
mSLAM-CTC (2B)	37.8	29.6	18.5	24.8
Maestro	**38.2**	31.3	18.4	25.2
Zero-Shot Whisper	36.2	**32.6**	**25.2**	**29.1**

Testset: CoVoST2

Github: https://github.com/facebookresearch/covost

3. Experiment

(4) Comparison with Human Performance



- Testset: Kincaid46 데이터셋에서 25개의 녹음을 선택
- A–D: 자동음성인식 상용 서비스
- E: 컴퓨터 전사 제공, 작업자가 전사 작성
- F–I: 전문 텍스트 변환 작업자가 작성한 전사
- Whisper는 인간과 유사한 수준을 보임.

4. Conclusion

- ✓ Limitations: 전혀 관련 없는 transcribe 생성, 장문 transcribe의 어려움, 반복 generation 문제 등
- ✓ Weakly supervised pre-training 을 통한 STT 성능 향상
- ✓ 68만 시간의 labeled audio data를 학습
- ✓ Text Standardization 제안
- ✓ 110개 언어지원 및 다양한 Task(번역, 언어감지, 음성전사 등) 가능

End of Document