# Robust Speech Recognition via Large-Scale Weak Supervision

# 목차

- Introduction
- Approach
- Experiments
- Analysis and ablations
- Limitations and Future Work
- Conclusion

# Introduction

- Wav2Vec 2.0
  - Encoder
    - Self-supervised learning
  - Decoder
    - 1,000,000 hours
    - Fine-tuning

# Introduction

- New Datasets
  - 680,000 hours of audio
  - 117,000 hours: 96 other languages
  - 125,000 hours: translation data
  - multilingual and multitask training

# Approach

- Data Processing
  - The raw text of transcripts
    - Without any significant standardization
  - Audio - Transcription
    - Environments, Recording setups, Speakers, and Languages
    - Not humangenerated transcripts
      - 기계가 만든 것을 학습시키면 성능 저하됨

# Approach

- Data Processing
  - Remove
    - 음성 신호만으로 알 수 없는 것
      - Ex) “…”, “?”, “!”, “,”
    - All-uppercase or all-lowercase transcript

# Approach

- Data Processing
  - Audio language detector
    - Fine-tuning
      - Model: CLD2
      - Dataset: VoxLingua107
    - 음성 언어와 전사 언어가 일치하면 데이터 셋에 추가
    - Fuzzy de-duping 사용
      - 전사 중복 및 자동 생성된 콘텐츠 양 줄이기 위함

# Approach

- Data Processing
  - Break audio file into 30-second segments
  - train on all audio(with no speech audio)

  - Training an initial model and remove low-quality data
    - High error rate
    - Data source size

# Approach

- Model

| | |
|---|---|
| Audio sampling rate | 16kHz |
| Channel | 80 |
| Audio feature normalization | -1 ~ 1 |
| Encoder | 2 convolution layer<br>Activation : GELU, GELU<br>Filter : 3, 3<br>Stride: 1, 2<br>Transformer |
| Decoder | Learned position embeddings and tied input-output token representation<br>Same with and number of transformer blocks |

# Approach

- Multitask Format
  - Ex)
    - 음성 감지
    - 화자 분리
    - 전사
    - 번역
    - 언어 식별

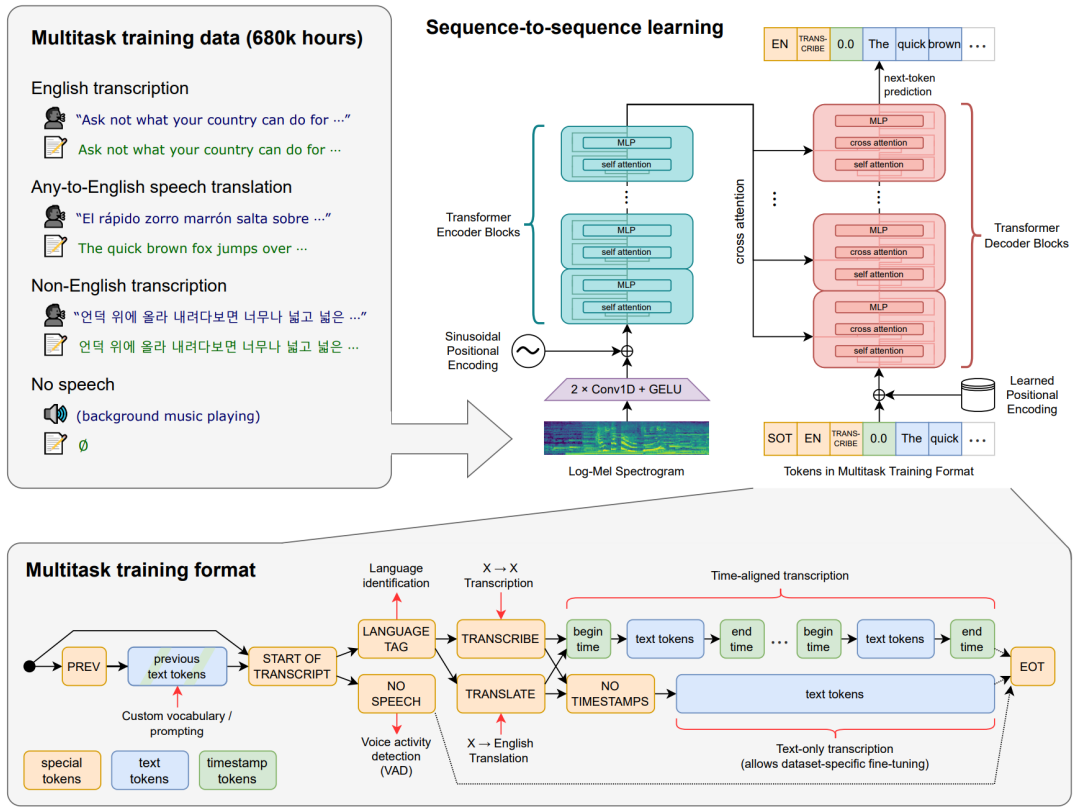


*Figure 1.* **Overview of our approach.** A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

# Approach

- Training Details

| | |
|---|---|
| Accelerator | FP16 |
| Optimizer | AdamW<br>Gradient norm clipping |
| Batch size | 256 |

| Model | Layers | Width | Heads | Parameters |
|---|---|---|---|---|
| Tiny | 4 | 384 | 6 | 39M |
| Base | 6 | 512 | 8 | 74M |
| Small | 12 | 768 | 12 | 244M |
| Medium | 24 | 1024 | 16 | 769M |
| Large | 32 | 1280 | 20 | 1550M |

*Table 1.* Architecture details of the Whisper model family.

# Experiments

- Zero-shot Evaluation
  - A single robust speech processing system that works reliably without the need for dataset specific fine-tuning

# Experiments

- Evaluation Metrics
  - Word Error Rate(WER)
    - Spelling and Phonetic word
      - 70% = 칠 십 퍼센트
    - 줄임말
      - I'm = I am

# Experiments

## English Speech Recognition

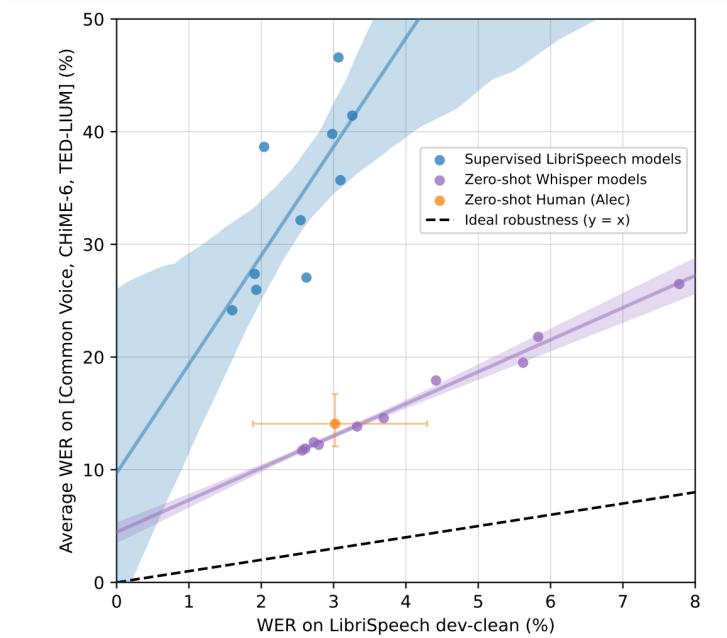


*Figure 2.* **Zero-shot Whisper models close the gap to human robustness.** Despite matching or outperforming a human on LibriSpeech dev-clean, supervised LibriSpeech models make roughly twice as many errors as a human on other datasets demonstrating their brittleness and lack of robustness. The estimated robustness frontier of zero-shot Whisper models, however, includes the 95% confidence interval for this particular human.

| Dataset | wav2vec 2.0 Large (no LM) | Whisper Large V2 | RER (%) |
|---|---|---|---|
| LibriSpeech Clean | **2.7** | **2.7** | 0.0 |
| Artie | 24.5 | **6.2** | 74.7 |
| Common Voice | 29.9 | **9.0** | 69.9 |
| Fleurs En | 14.6 | **4.4** | 69.9 |
| Tedlium | 10.5 | **4.0** | 61.9 |
| CHiME6 | 65.8 | **25.5** | 61.2 |
| VoxPopuli En | 17.9 | **7.3** | 59.2 |
| CORAAL | 35.6 | **16.2** | 54.5 |
| AMI IHM | 37.0 | **16.9** | 54.3 |
| Switchboard | 28.3 | **13.8** | 51.2 |
| CallHome | 34.8 | **17.6** | 49.4 |
| WSJ | 7.7 | **3.9** | 49.4 |
| AMI SDM1 | 67.6 | **36.4** | 46.2 |
| LibriSpeech Other | 6.2 | **5.2** | 16.1 |
| Average | 29.3 | **12.8** | 55.2 |

*Table 2.* **Detailed comparison of effective robustness across various datasets.** Although both models perform within 0.1% of each other on LibriSpeech, a zero-shot Whisper model performs much better on other datasets than expected for its LibriSpeech performance and makes 55.2% less errors on average. Results reported in word error rate (WER) for both models after applying our text normalizer.

# Experiments

## Multi-lingual Speech Recognition

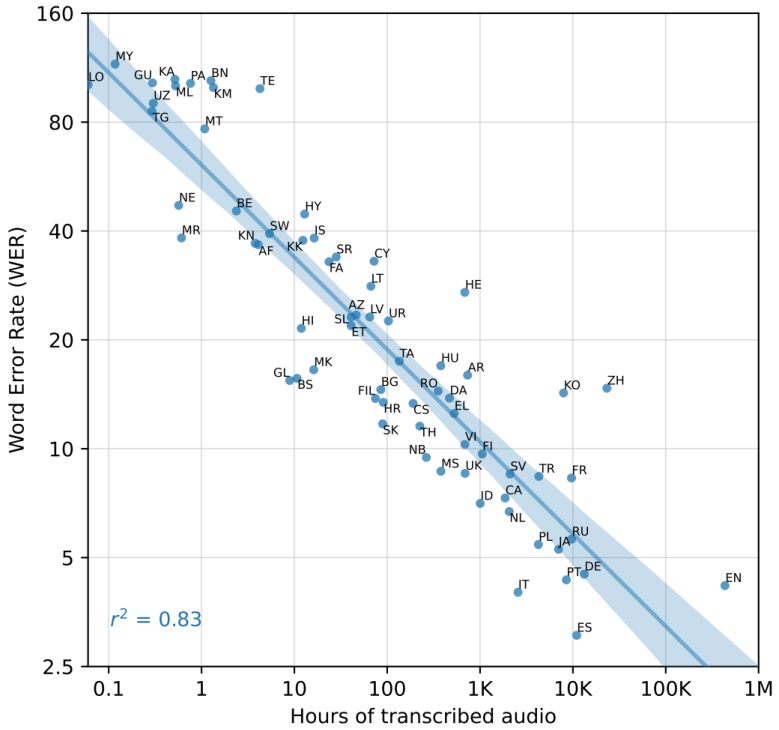


*Figure 3.* **Correlation of pre-training supervision amount with downstream speech recognition performance.** The amount of pre-training speech recognition data for a given language is very predictive of zero-shot performance on that language in Fleurs.

| Model | MLS | VoxPopuli |
|---|---|---|
| VP-10K + FT | - | 15.3 |
| XLS-R (1B) | 10.9 | 10.6 |
| mSLAM-CTC (2B) | 9.7 | 9.1 |
| Maestro | - | **8.1** |
| Zero-Shot Whisper | **7.3** | 13.6 |

*Table 3.* **Multilingual speech recognition performance.** Zero-shot Whisper improves performance on Multilingual LibriSpeech (MLS) but is still significantly behind both Maestro, XLS-R, and mSLAM on VoxPopuli.

# Experiments

- Translation

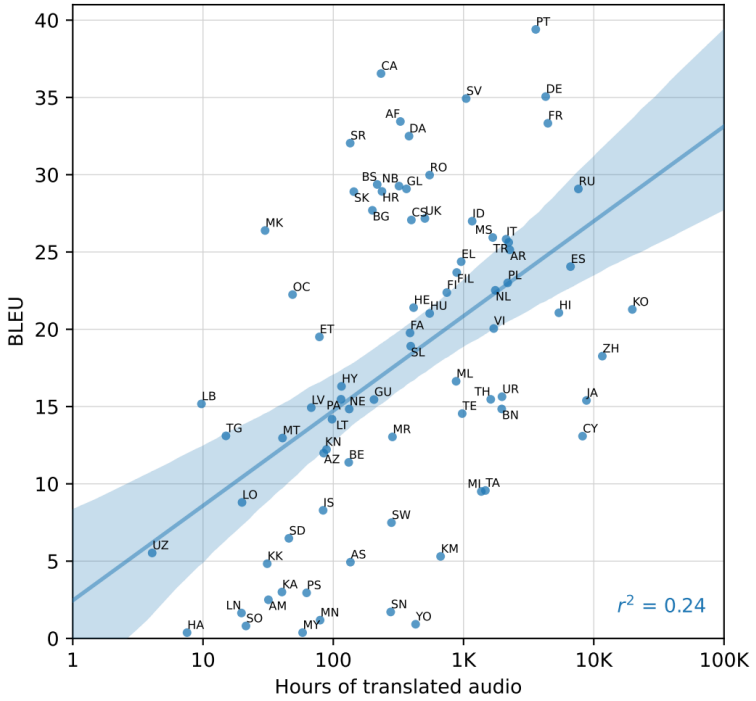


*Figure 4.* **Correlation of pre-training supervision amount with downstream translation performance.** The amount of pre-training translation data for a given language is only moderately predictive of Whisper's zero-shot performance on that language in Fleurs.

| 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** |

*Table 4.* **X→en Speech translation performance.** Zero-shot Whisper outperforms existing models on CoVoST2 in the overall, medium, and low resource settings but still moderately underperforms on high-resource languages compared to prior directly supervised work.

# Experiments

- Language Identification

| Language ID | Fleurs |
|---|---|
| w2v-bert-51 (0.6B) | 71.4 |
| mSLAM-CTC (2B) | **77.7** |
| Zero-shot Whisper | 64.5 |

*Table 5.* **Language identification performance.** Zero-shot Whisper's accuracy at language identification is not competitive with prior supervised results on Fleurs. This is partially due to Whisper being heavily penalized for having no training data for 20 of Fleurs languages.

# Experiments

- Robustness to Additive Noise

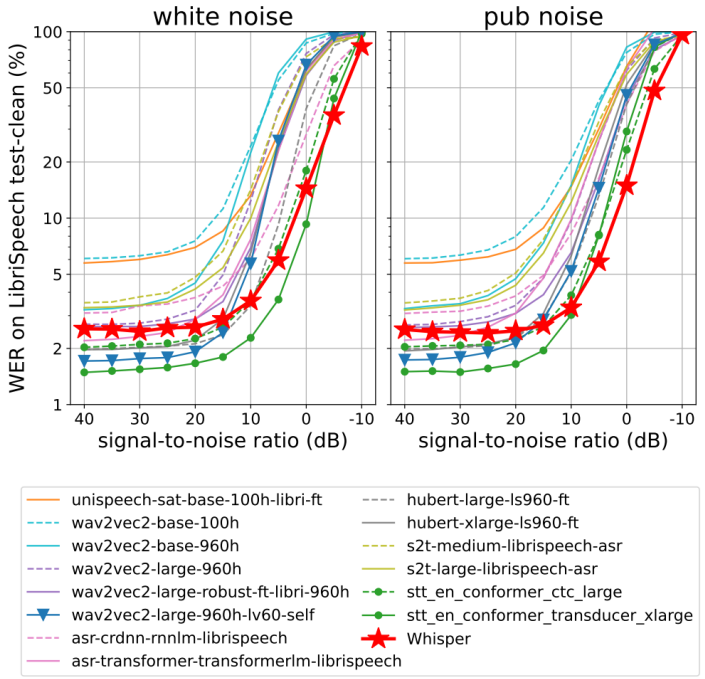


*Figure 5.* **WER on LibriSpeech test-clean as a function of SNR under additive white noise (left) and pub noise (right).** The accuracy of LibriSpeech-trained models degrade faster than the best Whisper model (★). NVIDIA STT models (●) perform best under low noise but are outperformed by Whisper under high noise (SNR $<$ 10 dB). The second-best model under low noise (▼) is fine-tuned on LibriSpeech only and degrades even more quickly.

# Experiments

- Long-form Transcription

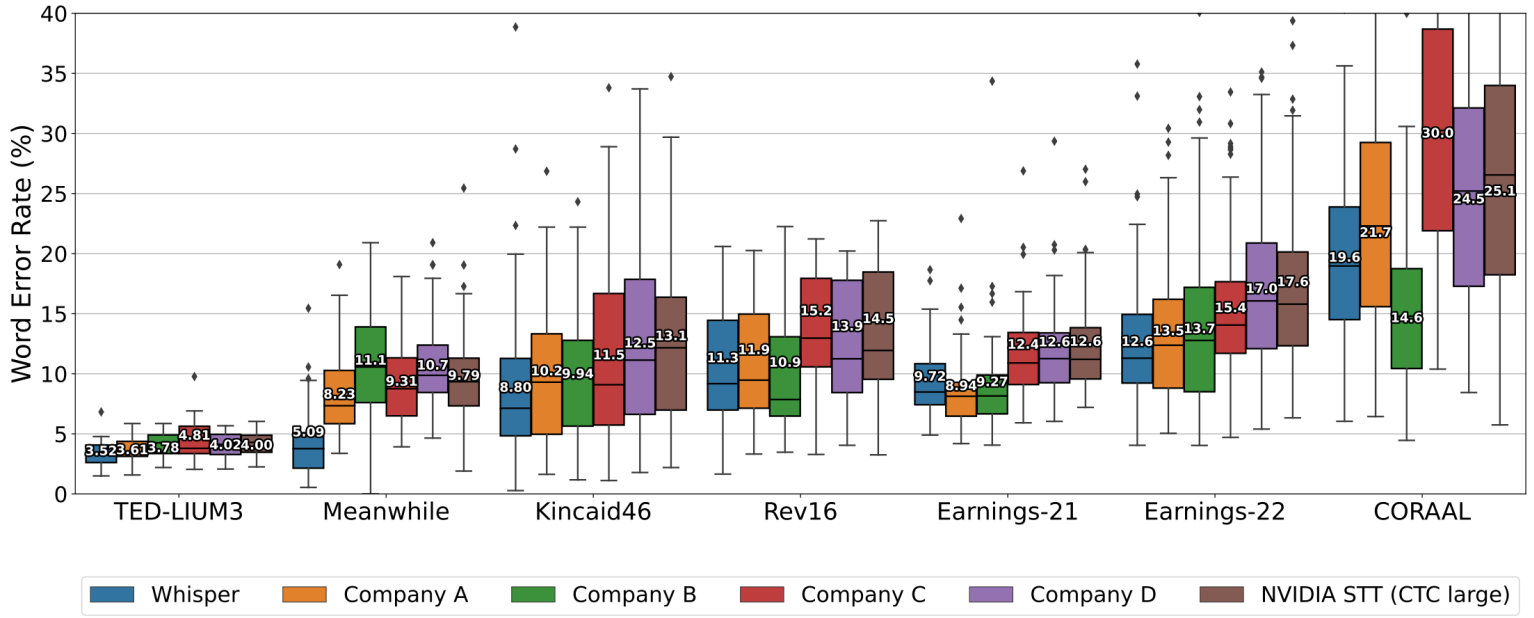


*Figure 6.* **Whisper is competitive with state-of-the-art commercial and open-source ASR systems in long-form transcription.** The distribution of word error rates from six ASR systems on seven long-form datasets are compared, where the input lengths range from a few minutes to a few hours. The boxes show the quartiles of per-example WERs, and the per-dataset aggregate WERs are annotated on each box. Our model outperforms the best open source model (NVIDIA STT) on all datasets, and in most cases, commercial ASR systems as well.

# Experiments

- Comparison with Human Performance

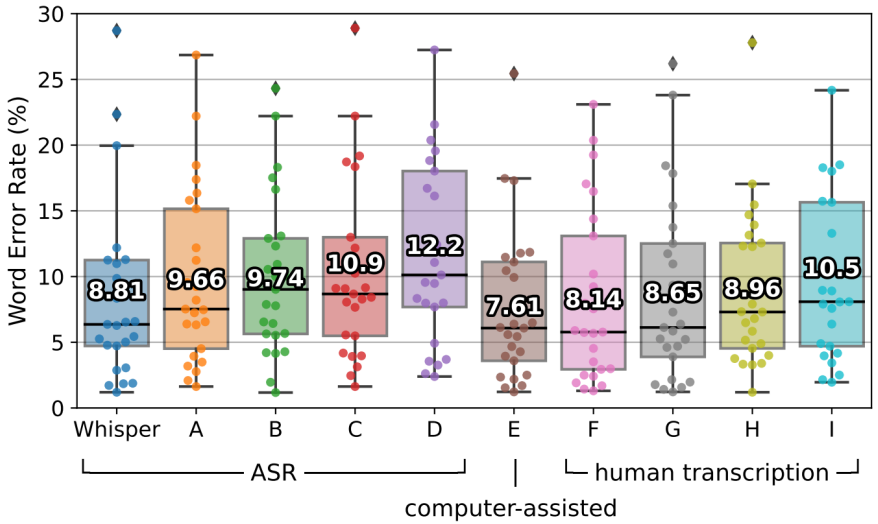


*Figure 7.* **Whisper's performance is close to that of professional human transcribers.** This plot shows the WER distributions of 25 recordings from the Kincaid46 dataset transcribed by Whisper, the same 4 commercial ASR systems from Figure 6 (A-D), one computer-assisted human transcription service (E) and 4 human transcription services (F-I). The box plot is superimposed with dots indicating the WERs on individual recordings, and the aggregate WER over the 25 recordings are annotated on each box.

# Analysis and ablations

- Model Scaling

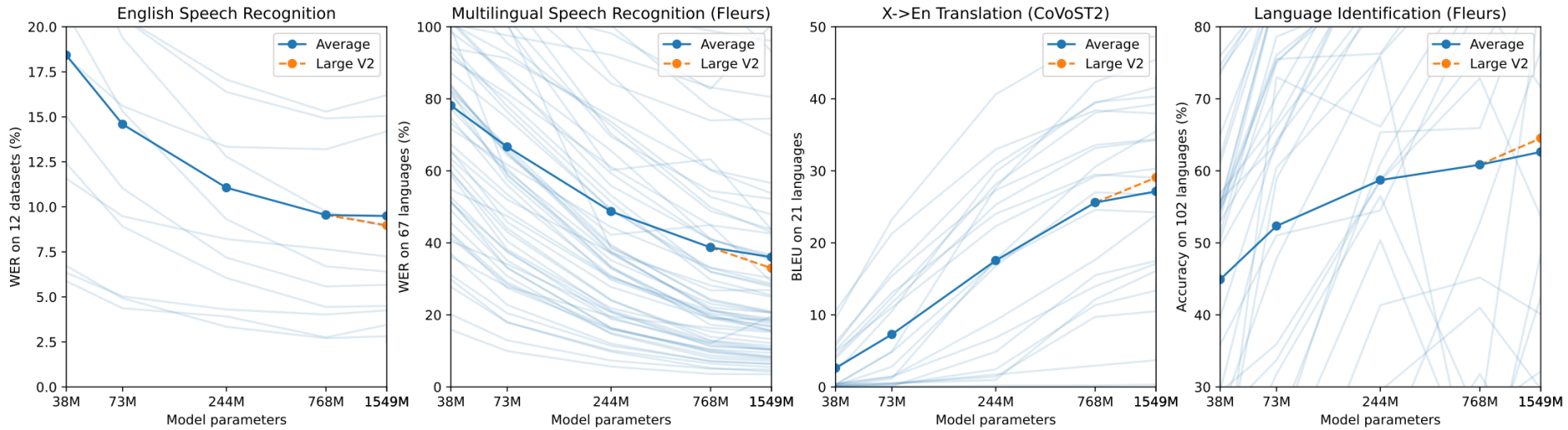


*Figure 8.* **Zero-shot Whisper performance scales reliably across tasks and languages with increasing model size.** Lightly shaded lines represent individual datasets or languages, showing that performance is more varied than the smooth trends in aggregate performance. Large V2 distinguished with a dashed orange line since it includes several changes that are not present for the smaller models in this analysis.

# Analysis and ablations

- Dataset Scaling

| Dataset size | English WER (↓) | Multilingual WER (↓) | X→En BLEU (↑) |
|---|---|---|---|
| 3405 | 30.5 | 92.4 | 0.2 |
| 6811 | 19.6 | 72.7 | 1.7 |
| 13621 | 14.4 | 56.6 | 7.9 |
| 27243 | 12.3 | 45.0 | 13.9 |
| 54486 | 10.9 | 36.4 | 19.2 |
| 681070 | **9.9** | **29.2** | **24.8** |

*Table 6.* **Performance improves with increasing dataset size.** English speech recognition performance refers to an average over 12 datasets while the Multilingual speech recognition reports performance on the overlapping subset of languages in Fleurs and X→en translation reports average BLEU on CoVoST2. Dataset size reported in hours.

# Analysis and ablations

- Multitask and Multilingual Transfer

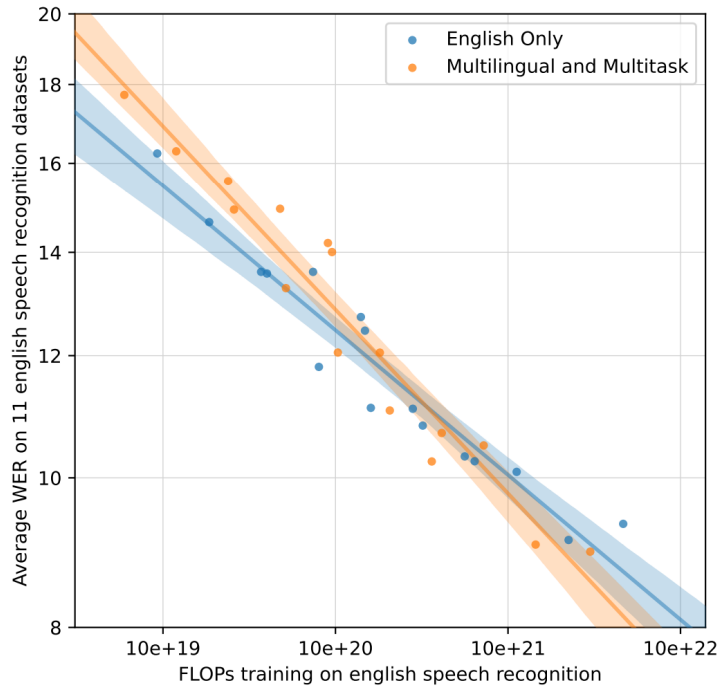


*Figure 9.* **Multitask and multilingual transfer improves with scale.** For small models, performance on English speech recognition degrades when trained jointly in a multitask and multilingual setup. However, multilingual and multitask models benefit more from scale and eventually outperform models trained on English data only. 95% bootstrap estimate confidence intervals are shown.

# Analysis and ablations

- Text Normalization

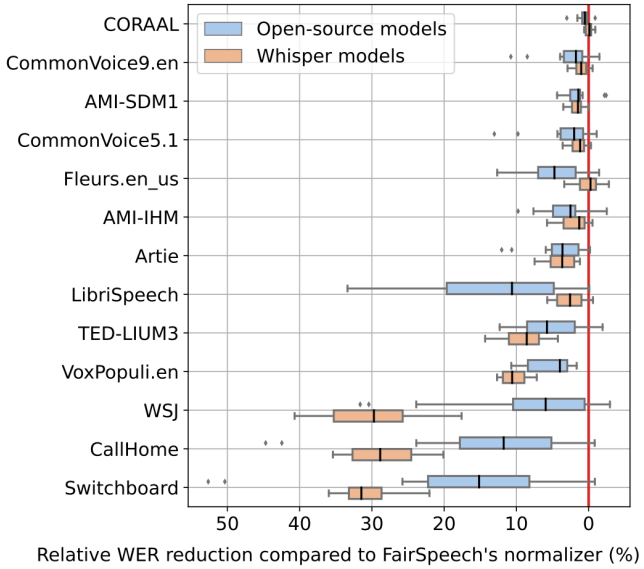


*Figure 10.* **On most datasets, our text normalizer has similar effect on reducing WERs between Whisper models and other open-source models, compared to FairSpeech's normalizer.** For each dataset, the boxplot shows the distribution of relative WER reduction across different models in our eval suite, showing that using our text normalizer generally results in lower WERs than FairSpeech's. On a few datasets our normalizer reduces WER significantly and more so for Whisper models, such as CallHome and Switchboard which have many contractions in the ground truth and WSJ which contains many numerical expressions.

# Analysis and ablations

- Strategies for Reliable Long-form Transcription

| | TED-LIUM3 | Meanwhile | Kincaid46 | Rev16 | Earnings-21 | Earnings-22 | CORAAL | Average |
|---|---|---|---|---|---|---|---|---|
| Greedy decoding only | 3.95 | 5.16 | 9.69 | 11.7 | 10.7 | 14.0 | 22.0 | 11.0 |
| + Beam search | 4.16 | 5.71 | 9.42 | 11.5 | 10.2 | 13.4 | 20.0 | 10.6 |
| + Temperature fallback | 4.16 | 5.71 | 9.42 | 11.5 | 10.2 | 13.4 | 20.0 | 10.6 |
| + Voice activity detection | 3.56 | **4.61** | 9.45 | 11.4 | 10.1 | 13.2 | 19.4 | 10.2 |
| + Previous text conditioning | **3.42** | 6.16 | 8.72 | **11.0** | **9.63** | 13.3 | **18.1** | 10.0 |
| + Initial timestamp constraint | 3.51 | 5.26 | **8.41** | 11.5 | 9.73 | **12.6** | 19.1 | 10.0 |

*Table 7.* Long-form transcription performance improves incrementally as additional decoding heuristics are employed. Details on each intervention are described in Section 4.5.

# Limitations and Future Work

- Improved decoding strategies
- Increase Training Data For Lower-Resource Languages
- Studying fine-tuning
- Studying the impact of Language Models on Robustness
- Adding Auxiliary Training Objectives

# Conclusion

- Supervised Learning
- Diverse supervised dataset
- Zero-shot transfer

# Link

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