LINGUISTIC BIAS IN AUTOMATIC SPEECH RECOGNITION FOR PEOPLE WHO STUTTER

Olin College of Engineering

Dongim Lee, Anna Du, Xavier Nishikawa, Anika Mahesh, Sreesanth Adelli, Troy Anderson, Kenneth Xiong

Public Interest Technology

dlee3@olin.edu

BACKGROUND

Keywords:
Automatic Speech Recognition (ASR), People Who Stutter (PWS), Word Error Rate (WER), Character Error Rate (CER)

ASR is widely used in tools like voice assistants and translators but often struggle to transcribe stuttered speech, impacting 80 million PWS. Our research question is: ***How can fine-tuning ASR models improve recognition of disfluent speech?***

We evaluated models on English and Mandarin stutters and fine-tuned them to reduce ***bias***[1] to make ASR technology more inclusive and accessible.

[1] Inequalities in transcription accuracy between fluent and stuttered speech

DATA & METHODS

Datasets

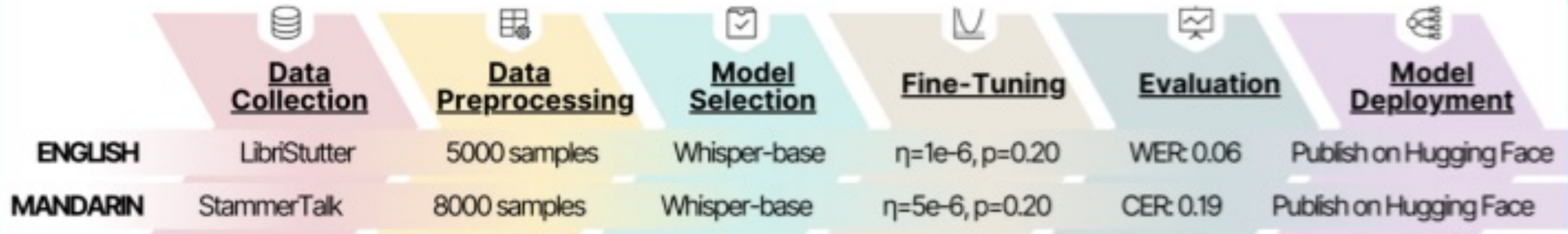
- **LibriStutter** [1]: 20 hours of **English** stuttered speech audio.
- **StammerTalk** [2]: 50 hours of **Mandarin** stuttered speech audio.

Stutter Types

1. **Word Repetition**: Repeating entire words (e.g., "I-I-I want").
2. **Sound Repetition**: Repeating single sounds (e.g., "b-b-b-ball").
3. **Phrase Repetition**: Repeating full phrases (e.g., "I like I like I like").
4. **Blocks**: Complete stoppage of speech, often with tension.
5. **Prolongation**: Stretching sounds out (e.g., "ssssee").
6. **Interjection**: Adding filler sounds or words (e.g., "um," "uh").

Model Training Pipeline

We fine-tuned **OpenAI's Whisper-base** ASR model [3] for each of the English and Mandarin dataset. Fine-tuning improves model performance by further training it on specific datasets to better handle stuttering.



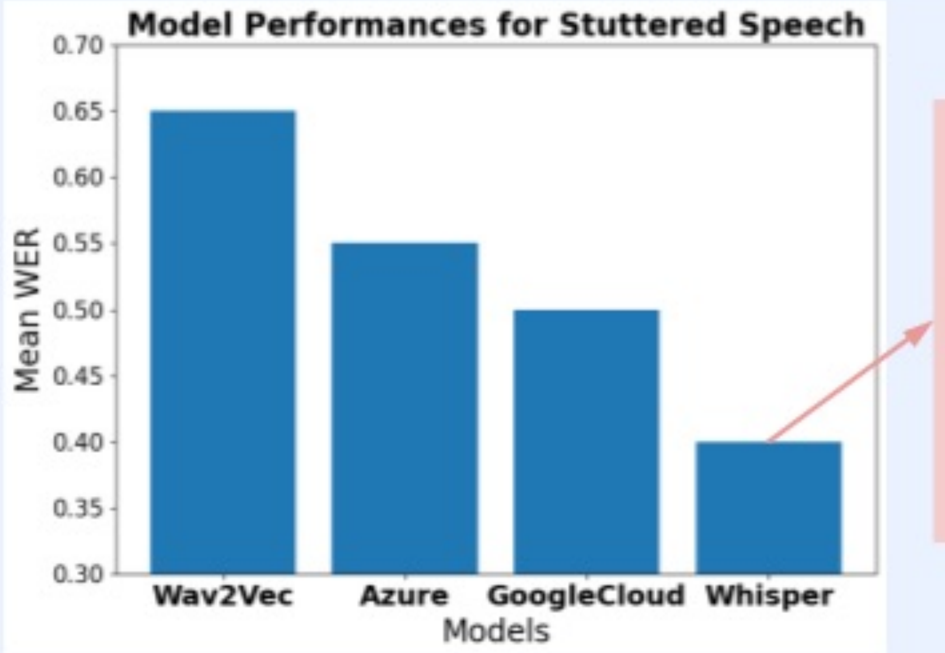
	Data Collection	Data Preprocessing	Model Selection	Fine-Tuning	Evaluation	Model Deployment
ENGLISH	LibriStutter	5000 samples	Whisper-base	η=1e-6, p=0.20	WER: 0.06	Publish on Hugging Face
MANDARIN	StammerTalk	8000 samples	Whisper-base	η=5e-6, p=0.20	CER: 0.19	Publish on Hugging Face

Figure 1. ASR fine-tuning workflow. Datasets were preprocessed, and Whisper-base was fine-tuned with learning rate (η) and dropout rate (p), evaluated, and published.

Evaluation Metrics

- **WER**(Word Error Rate) → Used for **English**'s word-based structure
- **CER**(Character Error Rate) → Used for **Mandarin**'s monosyllabic structure
- $WER(CER) = \frac{Insertions + Deletions + Substitutions}{Total\ Words(Characters)\ in\ Ground\ Truth}$

RESULTS



Whisper achieves the lowest WER and outperforms all evaluated ASR models on stuttered speech, which made it the ideal baseline for fine-tuning in our study.

Figure 2. Mean WERs of ASR models on stuttered speech.

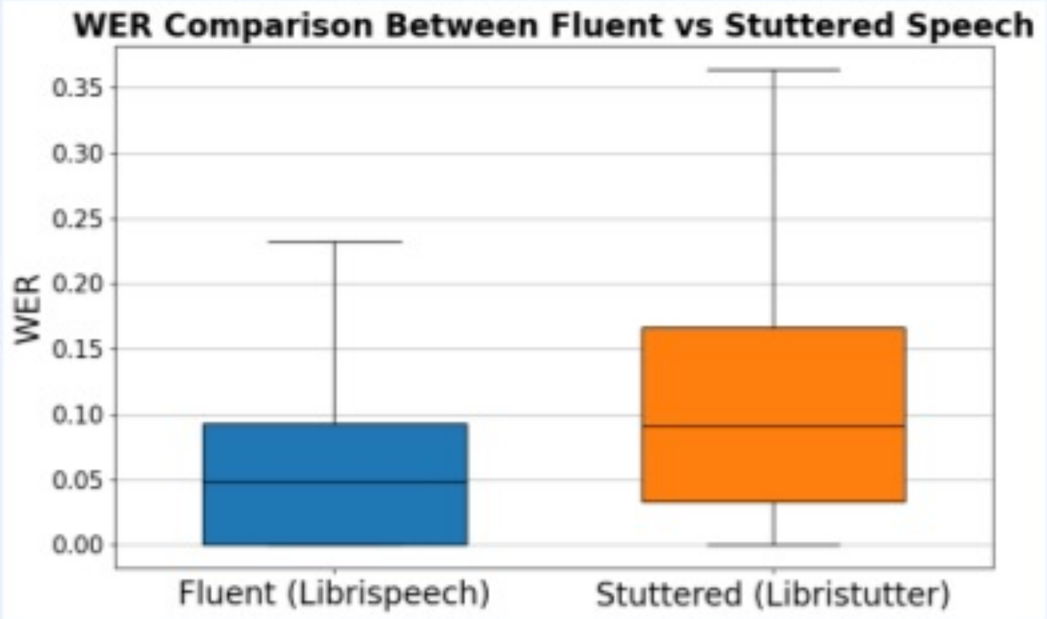
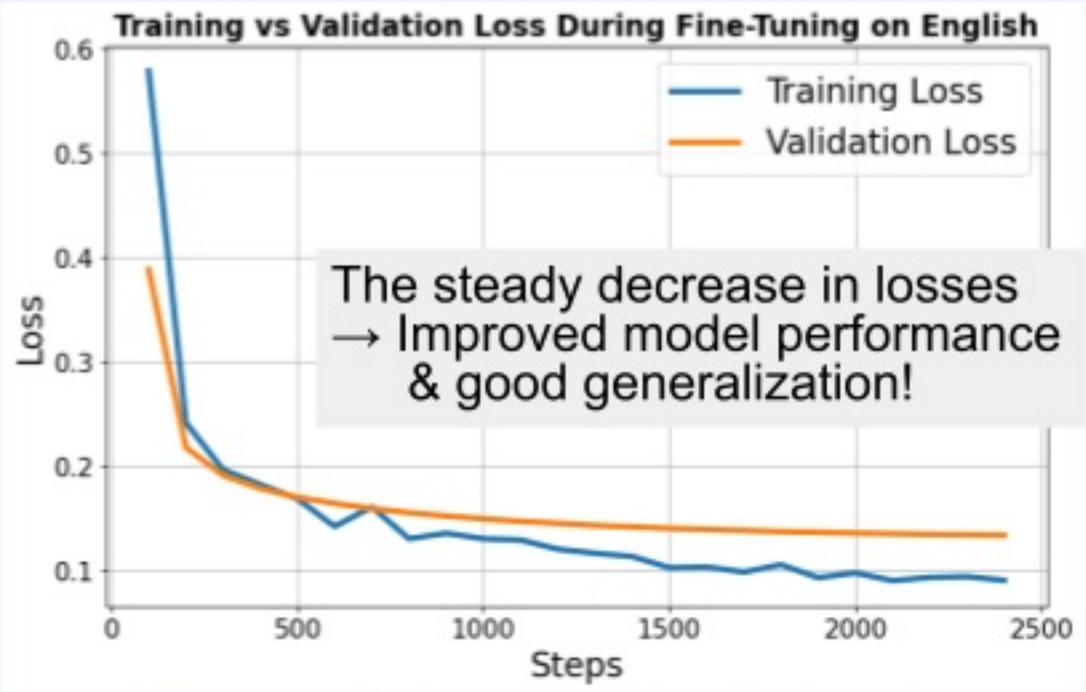


Figure 3. Baseline Whisper shows higher WER for stuttered speech, indicating difficulty with disfluency.

ENGLISH MODEL: WER 20.4% → 6.2% (-14.2%)



The steady decrease in losses → Improved model performance & good generalization!

Figure 4. Train/validation loss for English fine-tuning.

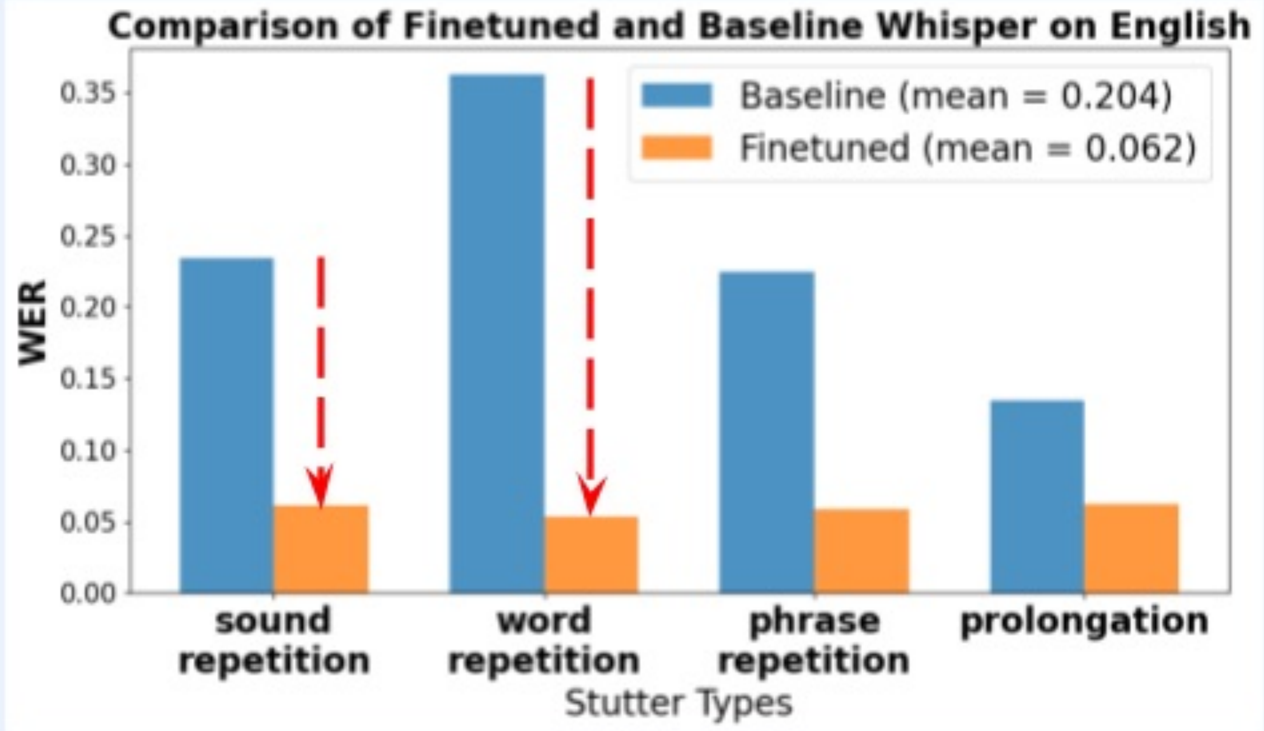


Figure 5. Compares baseline and fine-tuned English model's WER across stutter types.

MANDARIN MODEL: CER 66.4% → 19.0% (-47.4%)

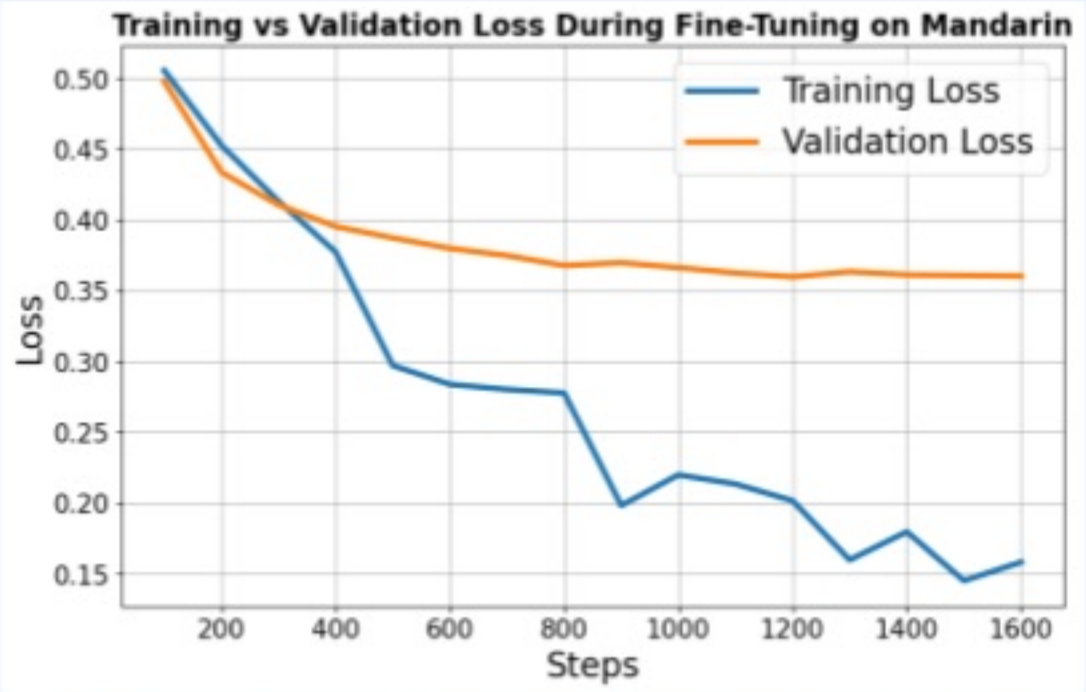


Figure 6. Train/validation loss for Mandarin fine-tuning.

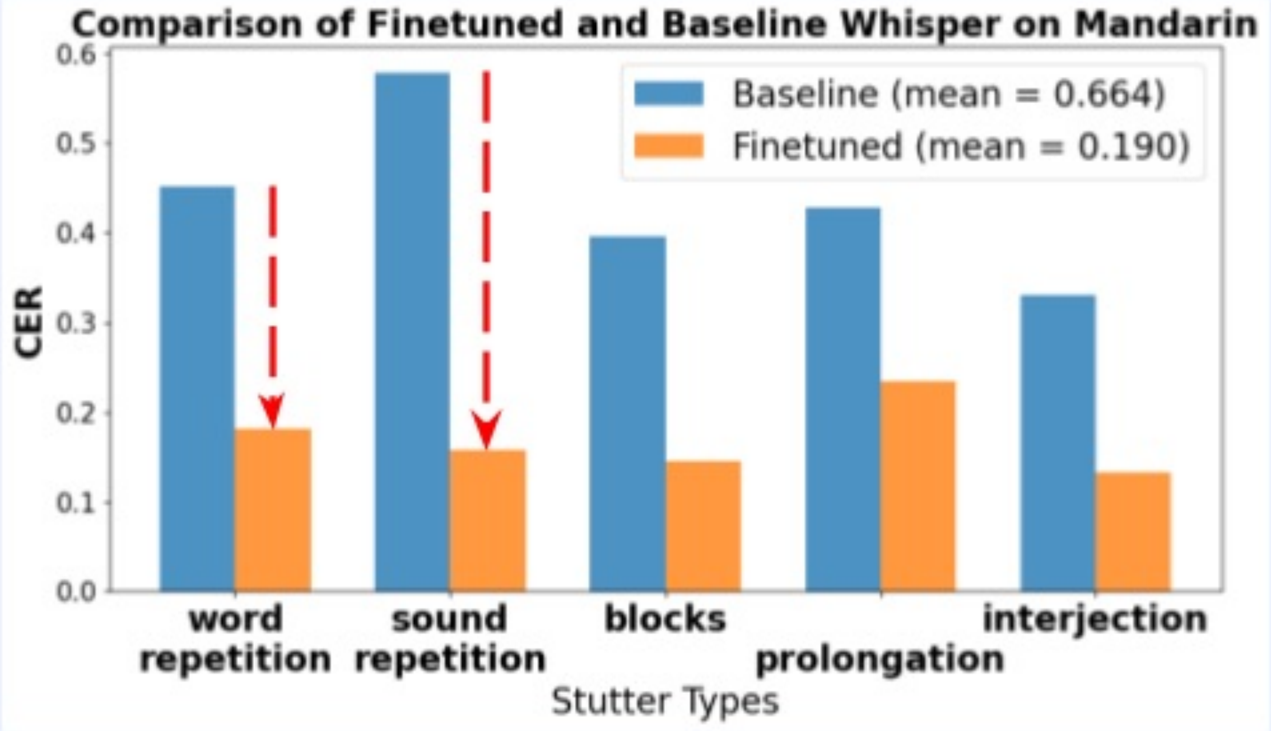


Figure 7. Compares baseline and fine-tuned Mandarin model's CER across stutter types.

Ground Truth	Whisper Base Transcription | WER 59%	Fine-Tuned Transcription | WER: 0%
… **doone** *(Word Repetition)* **…** to man that you're a liar he added **can you ever be** *(Phrase Repetition)* happy …	… **dune dune dune …** the cheerleyer he added **can you ever be can you ever be can you ever be** happy …	… **doone …** to man that you're a liar he added **can you ever be** happy …

Table 1. Example transcriptions from the baseline and fine-tuned models. The base model struggles with repetitions, while the fine-tuned model accurately removes repetitions, showing its effectiveness in handling disfluencies.

CONCLUSIONS

- **14.2% WER reduction** for the **English fine-tuned model** and **47.4% CER reduction** for the **Mandarin model**, proving their robustness and adaptability for stuttered speech.
  * Fine-tuned models are available on Hugging Face [4][5].
- While even large corporations like OpenAI have not fully addressed diverse speech patterns in ASR systems, our work sets the foundation for **inclusive AI that supports marginalized communities**.
- This research encourages ASR developers to prioritize inclusivity and build equitable AI for all, including 80M PWS.

KEY FINDINGS

- Fine-tuning significantly improved **sound/word repetitions** → Our models successfully address the **repetitive nature of stuttered speech**
- English Model: WER dropped to ~5% across all stutter types → Model performs stable on stuttered English
- Mandarin Model: **Prolongation** had the highest CER → **Mandarin is tonal language**. Elongated sounds could disrupt tonal patterns, causing inaccurate transcription.

→ **Linguistic factors** significantly impact ASR performance!

FUTURE WORK

- Fine-tune ASR models on diverse languages, dialects, and accents to expand inclusivity.
- Create publicly available stuttered speech datasets by collaborating with communities to address data scarcity.

CITATIONS

[1] Kourkounakis, T. (2021). *LibriStutter* (Version V1) [Dataset]. Borealis. https://doi.org/10.5683/SP3/NKVOGQ

[2] StammerTalk, AImpower, AIShell, Northwestern Polytechnical University, & Wenet Community. (2023). *StammerTalk-speech-70 Dataset* [Dataset]. Licensed under CC BY-NC 4.0.

[3] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision.* arXiv. https://arxiv.org/abs/2212.04356

[4] Fine-tuned Whisper English model. Available at: https://huggingface.co/dongim04/whisper-base-en

[5] Fine-tuned Whisper Mandarin model. Available at: https://huggingface.co/dongim04/whisper-base-zh