LINGUISTIC BIAS IN AUTOMATIC SPEECH RECOGNITION FOR PEOPLE WHO STUTTER

Dongim Lee, Anna Du, Xavier Nishikawa, Anika Mahesh, Sreesanth Adelli, Troy Anderson
- Olin College of Engineering
dlee3@olin.edu

BACKGROUND

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

ASR systems are widely used in technologies like voice assistants and translators. However, over 80 million PWS face significant challenges as ASR struggles to transcribe their speech accurately. This study explores: **How can fine-tuning ASR models improve recognition of disfluent speech?**

We evaluated ASR performance on stuttered speech with various stutter types in English and Mandarin, fine-tuning models to reduce bias i.e. inequalities in transcription accuracy between fluent and stuttered speech. By enhancing accuracy for stuttered speech, we aim to make ASR technology more inclusive and accessible.

DATA & METHODS

Datasets
To evaluate ASR models on stuttered speech, we utilized two datasets:

- LibriStutter [1]: 20 hours of English audio from 50 speakers, labeled with stutter types including sound repetition, word repetition, phrase repetition, and prolongation.
- StammerTalk [2]: 50 hours of Mandarin audio from 72 speakers, labeled with stutter types such as word repetition, sound repetition, blocks, prolongation, and interjection.

For training, the first 5000 audio samples from LibriStutter and the first 8000 samples from StammerTalk were used to fine-tune the models.

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 both datasets:

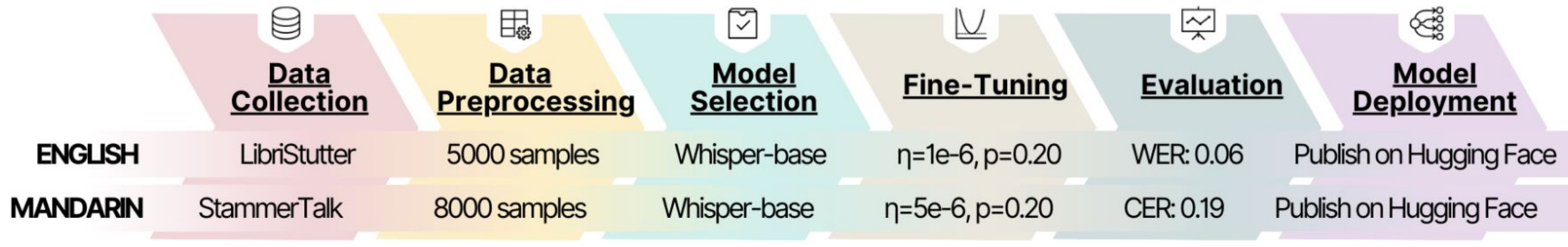
- English Model: Learning rate = 1e-6, Dropout rate = 0.20
- Mandarin Model: Learning rate = 5e-6, Dropout rate = 0.20

Fine-tuning involved adjusting the model's parameters to improve its transcription accuracy for stuttered speech patterns. Fine-tuning refers to refining a pre-existing ASR model by training it further on specific datasets to better handle unique challenges like stuttering.

Evaluation Metrics
Model performance was measured using:

- Word Error Rate (WER) for English:
  - $WER = \frac{Insertions+Deletions+Substitutions}{Total\ Words\ in\ Ground\ Truth}$
  - Measures errors at the word level, suitable for English's word-based structure.
- Character Error Rate (CER) for Mandarin:
  - $CER = \frac{Insertions+Deletions+Substitutions}{Total\ Charactors\ in\ Ground\ Truth}$
  - Measures errors at character level, suitable for Mandarin's monosyllabic structure.

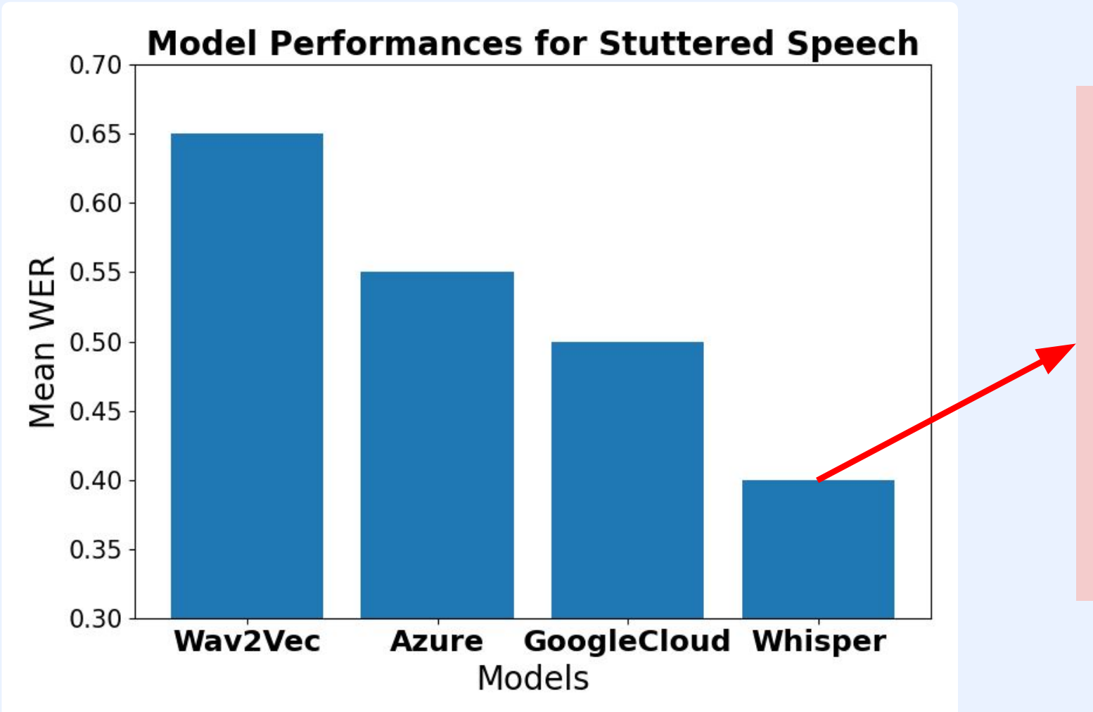
This ensures evaluations are accurate for each language.



	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 at 16,000 Hz. Whisper-base was fine-tuned with learning rate (η) and dropout rate (p), evaluated using WER and CER, and published on Hugging Face.

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 4 ASR models (Wav2Vec, Azure, Google Cloud Speech-to-Text, and Whisper) on stuttered speech are evaluated. Whisper has the lowest WER among all evaluated models, highlighting its robust performance as a baseline.

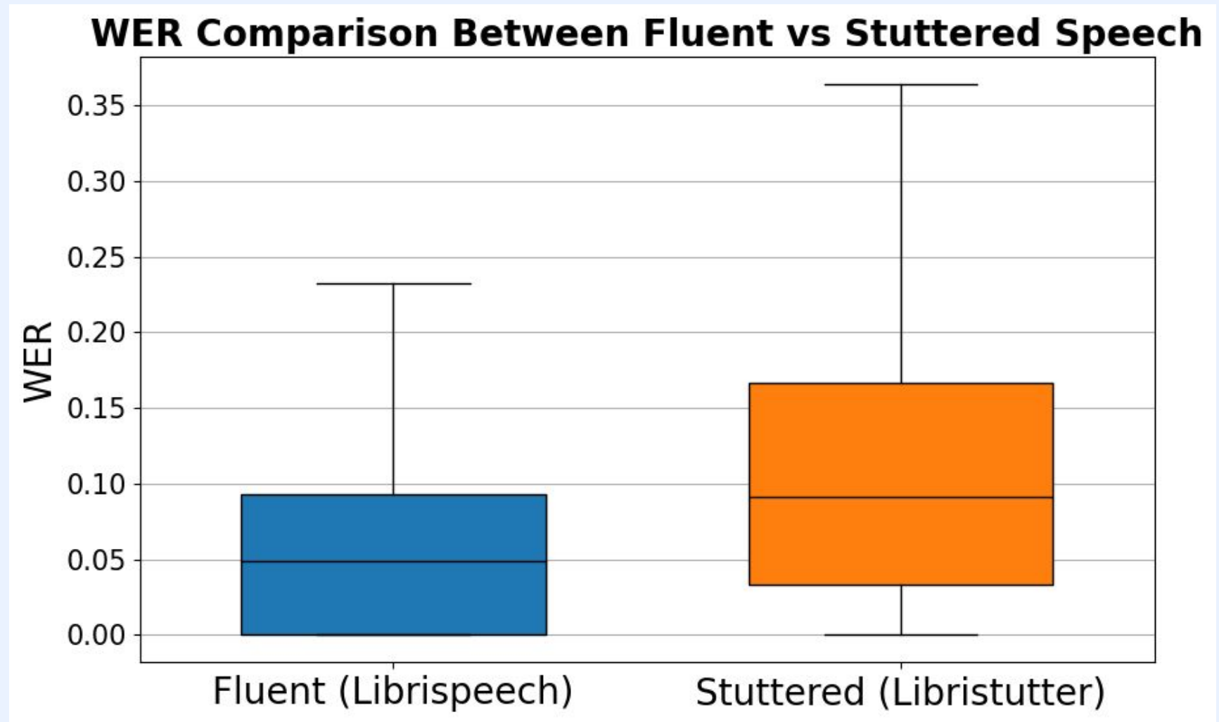


Figure 3. Shows the baseline Whisper's WER on fluent speech versus stuttered speech. The results show a significantly higher WER for stuttered speech, indicating model's struggle handling disfluent speech.

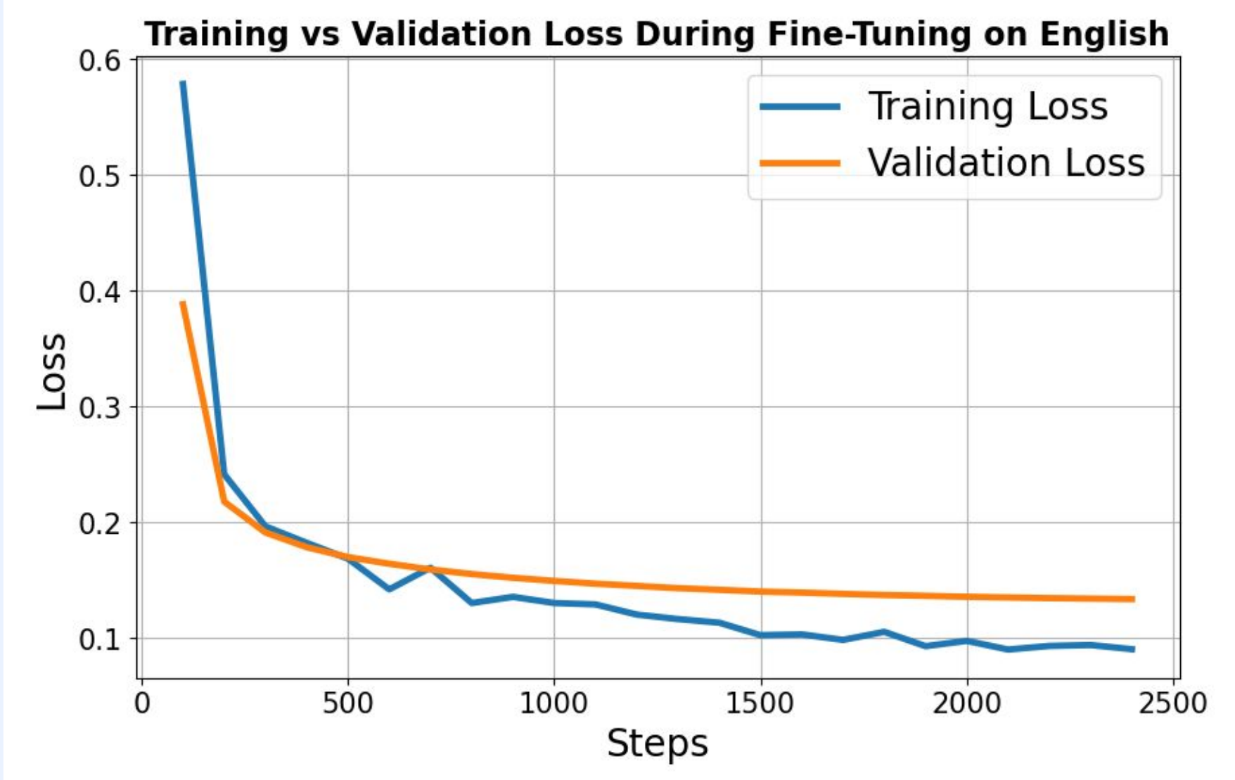


Figure 4. Training and validation loss curves during fine-tuning Whisper on English stuttering dataset. The steady decrease in losses indicates improved model performance without overfitting.

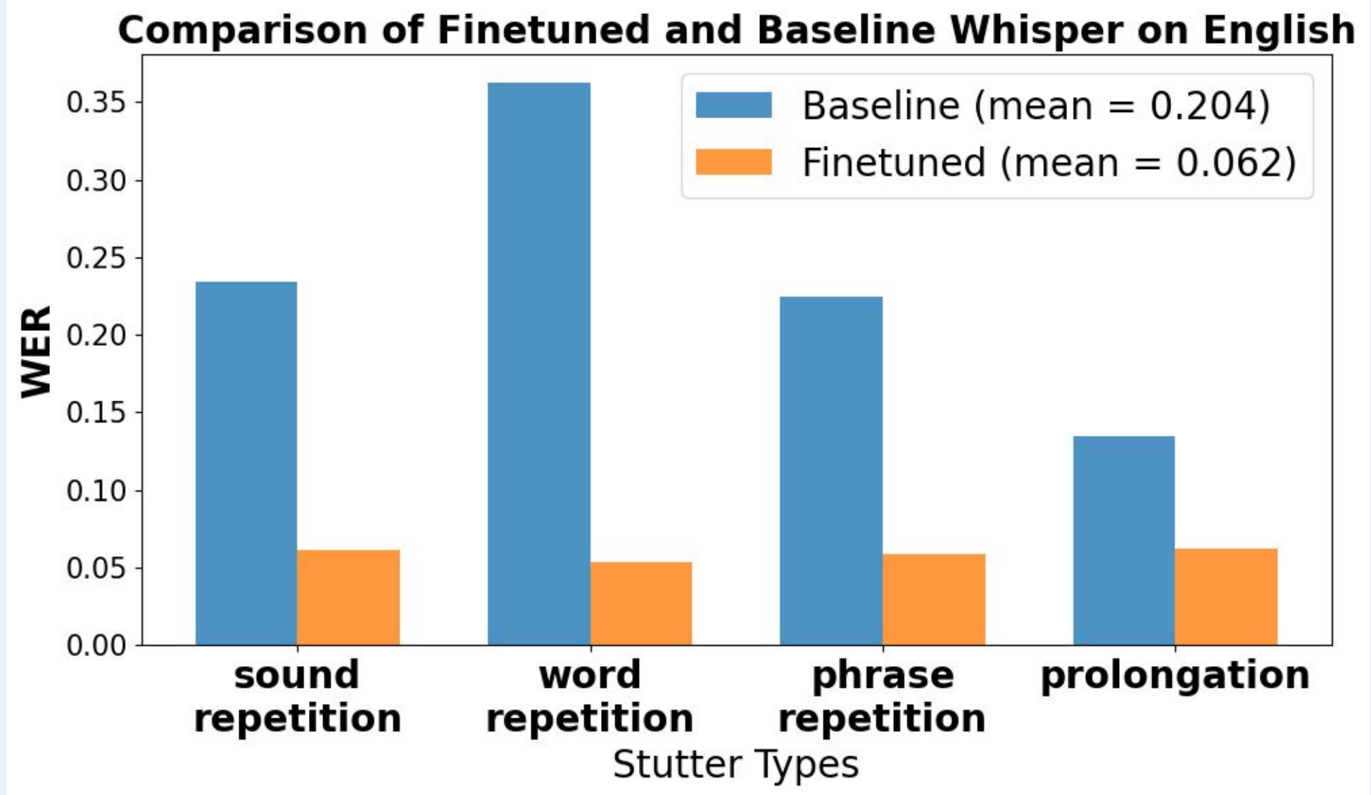


Figure 5. Compares the performance (WER) of the baseline and the fine-tuned Whisper model across various stutter types, including sound/word/phrase repetition and prolongation.

ENGLISH MODEL

Fine-tuning reduces the mean WER from **20.4%** to **6.2%**, significantly improving the recognition of stuttered English speech. While the baseline model struggles most with word repetition, the fine-tuned model achieves consistently low WER across all stutter types, demonstrating its robustness and adaptability.

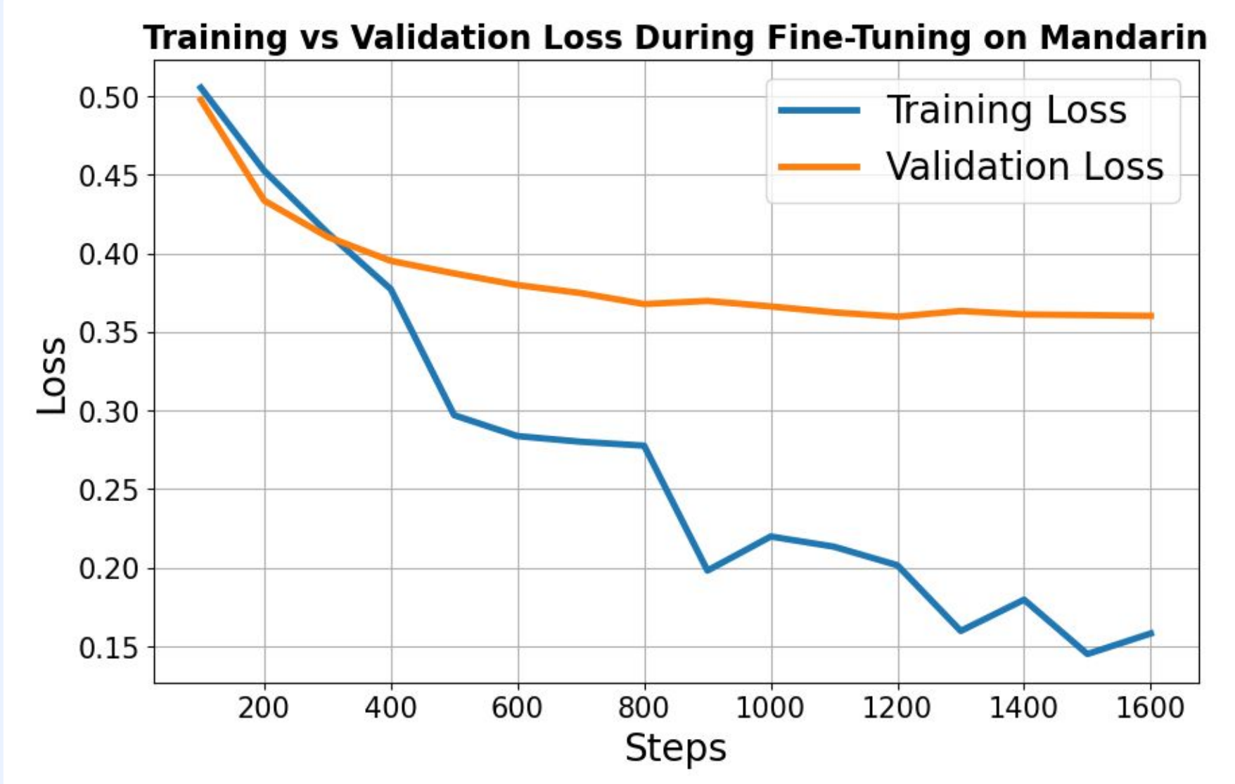


Figure 6. Training and validation loss curves during fine-tuning Whisper on Mandarin stuttering dataset.

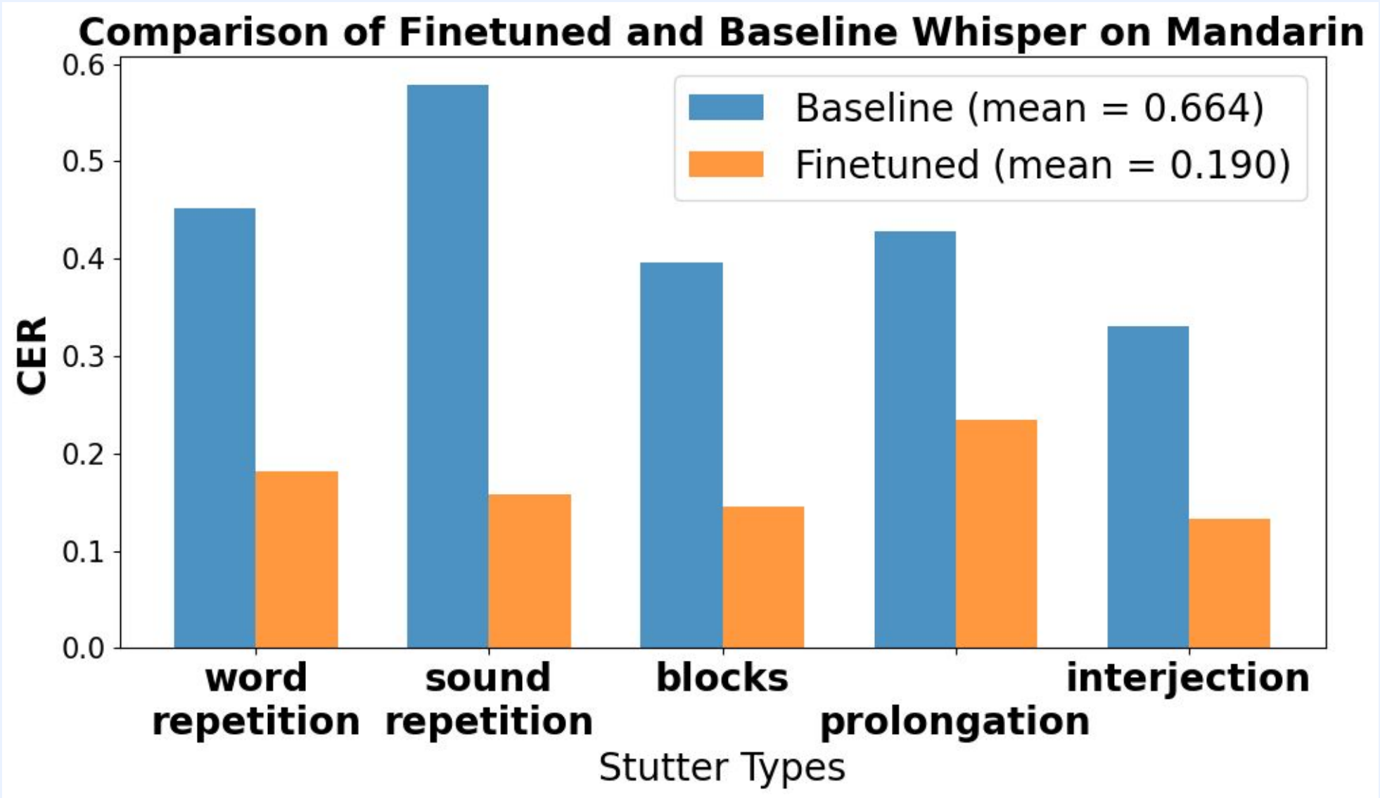


Figure 7. Compares baseline and fine-tuned Whisper model's performance (CER) across stutter types—word repetition, sound repetition, blocks, prolongation, and interjection.

MANDARIN MODEL

Fine-tuning reduces the mean CER from **66.4%** to **19.0%**, significantly improving the stuttered Mandarin speech recognition. While the base model struggles most with sound repetition, the fine-tuned model achieves consistently low CER across all stutter types, also demonstrating robustness and adaptability.

Ground Truth	**Whisper Base Transcription** | WER: 59%	**Fine-Tuned Transcription** | WER: 0%
said ronicky **doone (Word Repetition)** bill look me in the eye and tell me man to man that you're a liar he added **can you ever be (Phrase Repetition)** happy without her	said ronnyky **dune dune dune** bill look me in the eye and tell me man to man the cheerleyer he added **can you ever be can you ever be can you ever be** happy without her	said ronicky **doone** bill look me in the eye and tell me man to man that you're a liar he added **can you ever be** happy without her

Table 8. Example comparing baseline and fine-tuned model transcriptions with ground truth. The base model struggles with repeated words and phrases, while the fine-tuned model accurately removes repetitions, showing its effectiveness in handling disfluencies.

CONCLUSIONS

Validation was conducted using a subset of data from the LibriStutter dataset (English) and the StammerTalk dataset (Mandarin) for accurate comparisons between model transcriptions and ground truth. Our fine-tuned models achieved a **14.2% decrease in WER for English**, and a **47.4% decrease in CER for Mandarin** compared to the baseline model. These result highlight that our fine-tuning was extremely effective. The fine-tuned models are publicly available on Hugging Face [4][5] for anyone to use, enabling researchers, developers, and communities to build upon our work and further improve ASR systems for disfluent speech.

These results, achieved after six months of focused research and application, show a glaring oversight in ASR systems developed by large corporations including OpenAI's Whisper. Despite their immense resources, they failed to adequately account for diverse speech patterns. **Our model sets a precedent for making AI systems more inclusive from the ground up.** This work could serve as a catalyst for redefining industry standards, ensuring future ASR systems are built with a deeper understanding of diverse speech characteristics, ultimately creating more equitable AI solutions for all users, including the 80 million PWS.

KEY FINDINGS

- Sound/word repetitions were the worst performing stutter types for both English and Mandarin but showed the greatest improvement after fine-tuning, with error rates comparable to other stutter types. This demonstrates that the fine-tuned models effectively address the repetitive nature of stuttered speech.
- For the English fine-tuned model, the WER decreased to around 5% across all stutter types, showing that fine-tuning effectively bridges the performance gap for stuttered English speech.
- The Mandarin fine-tuned model showed more variability in error rates across different stutter types, with a mean CER of 19%. Prolongation had the highest error rate, likely due to Mandarin's tonal nature, where tones play a key role in distinguishing meaning. Elongated sounds can disrupt tonal patterns, introducing ambiguity and posing challenges for accurate transcription.
- The study highlights the importance of linguistic factors, such as tonal variations in Mandarin and repetitive patterns in stuttering, in ASR performance.

FUTURE WORK

- **Expanding Inclusivity**: Fine-tune speech recognition models for diverse languages, dialects, accents, and stutter types.
- **Addressing Dataset Scarcity**: Collaborate with institutions and communities to create publicly available stuttered speech datasets covering diverse patterns.
- **Enabling Public Access**: Provide fine-tuned models to the public through open-source platforms for practical use.

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