Reducing Bias in State-of-the-Art ASR Systems for Child Speech

Addressing Age and Gender Disparities through Transfer Learning Strategies

Franz Zeisler [f.a.zeisler@student.tudelft.nl]

Thesis committee: Zhengjun Yue, YuanYuan Zhang, Thomas Durieux

Background & Motivation

- **What is the problem?** Child speech presents unique challenges for ASR systems due to its distinctive acoustic and linguistic characteristics.
- **What are the challenges?**
  - Children's speech varies more in pitch, pronunciation, and articulation speed compared to adult speech [1].
  - This variability reduces accuracy in ASR systems designed for adult speech, impacting educational and communicative tools for children.
  - Biases in ASR systems result in lower recognition accuracy for children, with disparities across different age groups and genders.
- **What has been previously done?** Research has focused on specialised models using transfer learning and data augmentation. Despite improvements, data scarcity and biases remain significant challenges.

Research Question

How does fine-tuning affect recognition performance and biases across age and gender within child speech recognition using the Whisper model?

1. How effectively does the pre-trained Whisper model recognise child speech across age groups and genders?
2. What age and gender biases exist in the pre-trained Whisper model's recognition of child speech?
3. What changes occur in recognition performance after fine- tuning the Whisper model with child speech data?
4. How do age and gender biases in the Whisper model's recognition of child speech evolve following fine-tuning?

Overview Speech Corpora

Multiple speech corpora were used to enhance the validity of the research and test the generalisability of the results across different languages and speech types, including both spontaneous and non-spontaneous speech.

🇮🇸 Samromur Children [2] 🇳🇱 JASMIN-CGN [3] 🇩🇪 kidsTALC-v1 [4]

State-of-the-Art ASR Model - Whisper

Whisper is a state-of-the-art ASR model for speech recognition, developed by OpenAI in September 2022. It is available in various sizes, including tiny, base, small, medium, large-v1, large-v2, and large-v3. Trained on 680,000 hours of speech data, it efficiently handles **large-scale** data and employs **weak supervision**, allowing it to work with imperfect labels or transcripts. Additionally, it uses **zero-shot learning** to better adapt to new speech patterns and languages.

Metrics

Word Error Rate (WER): Percentage of words incorrectly predicted.

$$\text{WER} = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions}}{\text{Number of Words in Reference Transcript}} * 100\%$$

Bias: Disparities in WER across speaker groups (e.g., age, gender).

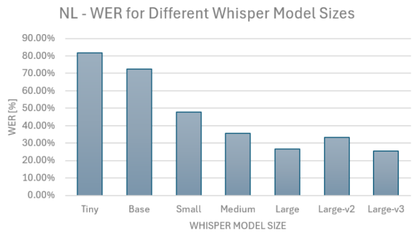
$$\text{Bias} = \text{WER}_{\text{group}} - \text{WER}_{\text{min}}$$

Overview Methodology

1. **Data Preprocessing:** The three speech corpora were preprocessed and split into training, validation and test sets.
2. **Benchmarking of Original Whisper Model:** The baseline performance of the unmodified Whisper model was established as a reference point for subsequent fine-tuning.
3. **Fine-Tuning of the Whisper Model:** Employed Low Rank Adaptation to fine-tune the pre-trained Whisper model using the child-specific datasets.
4. **Performance Evaluation:** The performance of the fine-tuned Whisper model was assessed again to assess improvements in recognition accuracy and reduction of age and gender biases.

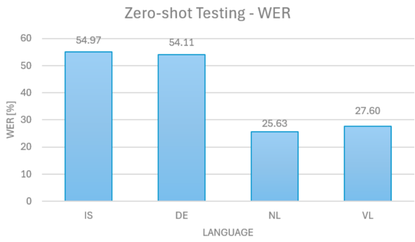
Model Size Selection

- Whisper has significantly higher WER for child speech compared to adult speech
- Large-v3 model selected for fine tuning since it performed best (lowest WERs and lowest bias)
- Large-v2 model performs poorly due to hallucinations



Pre-finetuning Results - WER

- Whisper training data:
  - German: 13,344 hours
  - Dutch: 2,077 hours
  - Icelandic: 16 hours
- Hypothesis: Best performance on German, followed by Dutch/Flemish, poorest on Icelandic.
- Factors affecting German ASR performance: spontaneous speech, age of children, over-trained on public broadcast subtitles



Pre-finetuning Results - Bias

- Bias rates were calculated for each specific age and gender demographic using the Word Error Rates (WERs).
- Gender bias: No consistent bias found across languages, indicating robust gender recognition.
- Age bias: Age biases are more pronounced than gender biases. Notably, biases decrease with increasing age groups, indicating Whisper's varied performance across different age demographics.

Gender	IS	DE	NL	VL
Female	0.00	6.27	6.52	0.00
Male	3.75	0.00	0.00	10.30

Age	IS	DE	NL	VL
3-5	N/A	29.18	N/A	N/A
6-8	14.66	30.06	15.41	19.55
9-11	9.98	0.00	9.54	0.00
12+	0.00	N/A	0.00	5.64

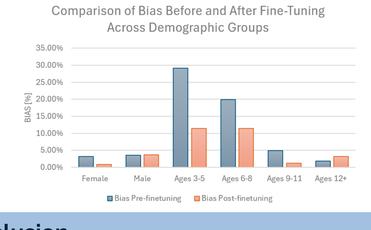
Post-finetuning Results - WER

- WER results improved significantly across all demographic groups.
- Average improvement of approximately 15%.
- Smallest improvement observed among teenagers.
- Average absolute WER is 37.08%.
- WER for current results is still much worse than WERs achieved for healthy adult speech.

Demographic Group	Change in WER
Female	-19.23%
Male	-13.23%
Ages 3-5	-16.16%
Ages 6-8	-19.11%
Ages 9-11	-13.58%
Ages 12+	-8.44%
AVG	**-14.96%**

Post-finetuning Results - Bias

- Gender bias reduced on average by 32.77%.
- Age bias reduced on average by 27.52%.
- For children aged 3-5 years, there was a relative WER reduction of 60.80%.
- For children aged 12 and older, the bias increased.



Conclusion

1. The pre-trained Whisper model initially struggled with recognising child speech accurately, especially in spontaneous contexts and with younger age groups (3-5 years), achieving a baseline WER of 40.58% using the large-v3 model across Icelandic, German, Dutch, and Flemish datasets.
2. Gender biases in the pre-trained Whisper model varied significantly across datasets, favouring female speakers in German and Dutch datasets and male speakers in Icelandic and Flemish datasets, with an average bias of 3.36%. Age biases were pronounced, averaging 13.97%, notably affecting younger age groups like those aged 3-5 years with a bias of 29.18%.
3. After fine-tuning with child speech data using Low-Rank Adaptation (LoRA), the Whisper model showed substantial improvement with an average WER reduction of 15.23%. The Dutch dataset demonstrated the most significant improvement with a WER reduction of 33.39%, achieving a post-finetuning WER of 17.69%.
4. Fine-tuning also led to notable reductions in gender biases by 32.77% on average across datasets and age biases by 27.52%. Particularly impressive was the reduction in bias for children aged 3-5 years, with a relative WER reduction of 60.80%, indicating improved recognition accuracy in this demographic.

References

[1] R. Paul and P. Flipsen. *Speech Sound Disorders in Children: In Honor of Lawrence D. Shriberg*. Plural Publishing, Incorporated, 2009. ISBN: 9781597567688. URL: https://books.google.nl/books?id=j1sOBwAAQBAJ.

[2] C. Mena et al. "Samromur Children: An Icelandic Speech Corpus". In: *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. Licensed under CC-BY-NC-4.0. European Language Resources Association (ELRA). Marseille, 2022, pp. 995–1002.

[3] C. Cucchiarini et al. "JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives in the Human-Machine Interaction Modality". In: *Proceedings of the Conference*. Available from: https://www.researchgate.net/publication/237245075_JASMIN-CGN_Extension_of_the_Spoken_Dutch_Corpus_with_Speech_of_Elderly_People_Children_and_Non-natives_in_the_Human-Machine_Interaction_Modality [accessed May 14, 2024]. 2015.

[4] L. Rumberg et al. "kidsTALC: A Corpus of 3- to 11-year-old German Children's Connected Natural Speech". In: *Journal Name* (2019). Available from: email:kidstalc@tnt.uni-hannover.de.