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

Addressing Age and Gender Disparities through Transfer Learning Strategies

Franz Zeisler[1]
Supervisor(s): Zhengjun Yue[1]**, YuanYuan Zhang**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2024

Name of the student: Franz Zeisler
Final project course: CSE3000 Research Project
Thesis committee: Zhengjun Yue, YuanYuan Zhang, Thomas Durieux

Abstract

Automatic Speech Recognition (ASR) systems have transformed human-machine interaction, yet they often struggle with child speech due to the unique vocal characteristics. This thesis investigates age and gender biases, focusing on enhancing the performance of state-of-the-art ASR model Whisper on child speech. Initial experiments reveal significant disparities in recognition accuracy across age groups and genders within child speech, highlighting the critical need for targeted improvements. The study uses Low-Rank Adaptation (LoRA) to finetune the model using four child-specific datasets, aiming to simultaneously enhance recognition performance and mitigate biases. Results demonstrate substantial reductions in Word Error Rates (WER) and biases after finetuning, showcasing the effectiveness of transfer learning in addressing demographic inequality. Gender biases decreased by 32.77% relative to their initial values, and age biases also improved, with a relative decrease of 27.52% after finetuning. This research showcases the potential of tailored approaches to advance ASR technology for low-resource user demographics, with implications for improving educational and assistive technologies.

Index Terms: Automatic Speech Recognition, Child speech, Whisper ASR model, Age and gender biases, Low-Rank Adaptation, Transfer learning, Demographic disparities

1. Introduction

Speech serves as the primary mode of human communication, providing an efficient and natural way to interact with both other humans and machines. With the rise of innovative AI technology for speech and language processing also came a large development of ASR systems. These systems are capable of transforming spoken language into written text whilst understanding context and nuances [1]. ASR systems are increasingly prevalent in various applications, such as search engines, voice assistants, educational tools, and accessibility services [2].

However, despite the advancements, most research efforts target ASR systems for healthy adult speakers [2]. This focus has left a noticeable gap in ASR capabilities when it comes to recognising child speech, which differs significantly from adult speech in pitch variability, pronunciation, and articulation speed [3]. These differences contribute to significantly lower performance compared to adults, with Word Error Rates (WER) being up to five times worse [4]. This limitation restricts the effectiveness of educational tools and assistive technologies for children [5, 6, 7].

Efforts to address these challenges have involved adapting models developed for adult speech to child speech using techniques such as data augmentation [8, 9] or transfer learning [10, 11, 12]. Recent experiments have enhanced ASR for child speech significantly through finetuning on child speech datasets. For instance, Southwell et al. achieved a relative reduction of 38% in WER using OpenAI's Whisper model on the MyST dataset [7], and Jain et al. reported absolute reductions in the range of 7% to 43% with low-resource finetuning techniques on other child datasets [13].

Despite these improvements, significant disparities in ASR performance persist across different age groups and genders. While adults and teenagers generally achieve higher accuracy in speech recognition, children and elderly speakers often experience notably lower performance levels [14, 15, 16]. Moreover, cross-linguistic investigations have revealed biases in speech recognition based on gender. Some studies report better recognition for male speakers [15, 17], while others found biases favouring female speakers [16, 18, 19], and still other studies show no gender-based differences [20].

Recognition bias in speech technology can result from several factors, including biased transcriptions, dialectal variations, under-representation of specific speaker groups in training data, equipment discrepancies, and intra-group variability in pronunciation and language use [21]. In particular, child speech recognition faces unique biases due to factors such as limited annotated datasets, developmental variations among young speakers, and inadequate diversity in training data [21]. These biases warrant the need for further research into methods that can mitigate age and gender biases within state-of-the-art ASR models like Whisper [22].

However, research is lacking on age and gender biases specific to child speech recognition within state-of-the-art ASR models, and methods to reduce these biases. This research aims to address the current research gap by initially exploring age and gender biases inherent in child speech recognition in the state-of-the-art ASR model Whisper. Subsequently, it seeks to determine whether finetuning, which has shown promising results in improving recognition performance, can also help mitigate these biases. By analysing the intertwined aspects of bias and recognition performance, this study aims to answer the research question: "**How does finetuning affect recognition performance and biases across age and gender within child speech recognition using the Whisper model?**" To achieve these objectives, the study addresses the following research questions sequentially:

1. How effectively does the pre-trained Whisper model recognise child speech across different 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 finetuning 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 finetuning?

The structure of the paper follows a systematic approach to investigate the enhancement of ASR systems for child speech. Section 2 introduces the Whisper ASR model and outlines the high-level methodology focusing on finetuning along with the performance metrics. Section 3 examines the three datasets used, detailing their division into test, training, and validation sets, and preprocessing. In Section 4, the experimental setup for the research is described, and relevant hyperparameters are introduced. The results obtained from benchmarking and post-finetuning evaluations are then presented and discussed in Section 5. Section 6 summaries the findings and suggests future research directions. Additionally, some ethical considerations and reproducibility are covered in Section 7.

2. Methodology

This research investigates how finetuning affects recognition performance and biases across age and gender in child speech recognition using the Whisper model. Subsection 2.1 intro-

duces the selected state-of-the-art ASR model, Whisper, providing an overview of its architecture and training process. The evaluation metrics used to assess the Whisper model's baseline performance, including Word Error Rate and bias calculations, are then discussed in Subsection 2.2. Finally, Subsection 2.3 details the chosen finetuning approach, specifically focusing on the Low-Rank Adaptation (LoRA) method.

2.1. State-of-the-Art ASR: Whisper Model

Whisper is an ASR model introduced by OpenAI in September 2022 [22]. Inspired by recent advancements in computer vision and natural language processing, it takes an approach that scales weakly supervised datasets in order to achieve robust and generalised models rather than relying heavily on traditional supervised training methods.

This approach has enabled Whisper to make significant advancements in ASR, outperforming established models such as Kaldi, DeepSpeech, SpeechBrain, and Wav2Vec 2.0. On the LibriSpeech dataset [23], Whisper achieved a WER of 5.2% [22], compared to DeepSpeech's 12.69% [24], Kaldi's 6.2% [25], SpeechBrain's 5.77% [26], and Wav2Vec2.0's 3.3% [27]. On the Common Voice corpus [28], Whisper achieved a WER of 9.0% [22], while DeepSpeech had 43.82% [29], Kaldi had 4.44% [25], SpeechBrain had 15.58% [26], and Wav2Vec 2.0 had 16.1% [30]. Whisper stands out as the best-performing model overall, particularly considering that Kaldi's dataset overlaps with Common Voice.

Whisper uses an encoder-decoder transformer architecture, a scalable model that performs various tasks. Interestingly, Whisper's training process is multitask: it takes on several aspects of speech processing in a single model, such as multilingual speech recognition, spoken language identification, speech translation, and voice activity detection. The model has been trained on a diverse dataset comprising 680,000 hours of labelled audio data, enabling it to generalise effectively across various types of speech in a zero-shot transfer setting. This includes 117,000 hours spanning 96 languages, 125,000 hours of translation data from various languages to English, and the remainder consists of English speech data.

The Whisper model family comprises variants with different numbers of parameters: tiny (39M), base (74M), small (244M), medium (769M), and large (1550M) [22]. Additionally, within the large variant, there are three versions: large-v1, large-v2, and large-v3. The large-v2 model, released in December 2022, shares the same size as the original large-v1 model but underwent training for 2.5 times more epochs [31]. In November 2023, the large-v3 model was introduced, maintaining the same architecture as its predecessors but trained on 1 million hours of weakly labelled audio and 4 million hours of pseudo-labelled audio from large-v2 over 2.0 epochs. This resulted in a further WER reduction ranging from 10-20% compared to large-v2 [32].

2.2. Evaluation Metrics

The baseline performance of the Whisper model on the datasets will be evaluated using WER as well as bias. WER, representing the percentage of words incorrectly predicted, serves as an accuracy benchmark and is calculated using:

$$WER = \frac{S + I + D}{N} * 100\%$$

where S represents the number of substitutions, I represents the number of insertions, D represents the number of deletions, and N represents the total number of words.[1]

For bias evaluation, the paper adopts the approach proposed by S. Feng et al. [21]. In this context, bias refers to the difference in WER across various speaker groups within each assessed dimension. This bias calculation can be represented by the following formula:

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

where $\text{WER}_{\text{group}}$ represents the WER of each speaker group in a dimension, and WER_{min} represents the lowest WER among all speaker groups within that dimension. This calculation allows for the assessment of demographic performance disparities across age and gender groups.[2]

This initial benchmarking will establish the Whisper model's starting accuracy and bias, providing a baseline for measuring performance improvements through transfer learning.

2.3. Finetuning Approach

This research evaluates and enhances Whisper's performance in recognising child speech. First, Whisper's baseline performance is established using WER and bias metrics across demographic groups. Next, the models are finetuned with child-specific datasets using transfer learning to improve recognition of child speech. Finally, the finetuned models are re-evaluated with the same metrics to assess the effectiveness of transfer learning in reducing biases and enhancing accuracy.

In machine learning, many approaches involve extensive pre-training on broad-domain data. However, as models grow larger, fully retraining all parameters becomes increasingly impractical and costly. This challenge is particularly significant for Whisper models, some of which comprise up to 1.55 billion parameters. A solution to this problem is provided by Low-Rank Adaptation (LoRA), a method for parameter-efficient finetuning [33]. By "freezing pre-trained model weights and injecting trainable rank decomposition matrices into each layer of the Transformer architecture" [33, p.1], LoRA drastically reduces the number of trainable parameters. Utilising the Hugging Face Parameter-Efficient-Finetuning (PEFT) library [34], LoRA enables efficient adaptation of the Whisper model to the nuances of child speech, facilitating more computationally and storage-efficient computations.

3. Datasets

To ensure the robustness and generalisability of the results, three diverse child speech corpora were selected: an Icelandic

[1] For instance, if a reference transcription contains 100 words, and the ASR system includes 5 substitutions, 3 insertions, and 2 deletions, the WER would be calculated as $\frac{5+3+2}{100} = 10\%$

[2] For instance, if the WER for male speakers is 60% and the WER for female speakers is 52%, then the male bias would be $60\% - 52\% = 8\%$ and the female bias would be $52\% - 52\% = 0\%$.

speech corpus, a German speech corpus, and a combined Dutch and Flemish corpus. These are introduced in Subsection 3.1, Subsection 3.2 and Subsection 3.3 respectively. Initial preprocessing was required for each dataset to prepare the speech corpora in a format suitable for training and finetuning the Whisper model. The transcripts were cleaned by removing non-verbal parts, annotations (e.g., unintelligible parts, overlapping parts, phonological/lexical errors), punctuation, and converting all text to lowercase. An overview of the processed corpora are provided in Subsection 3.4.

3.1. Samromur Children

This Icelandic speech corpus [35], aimed at ASR, contains 131 hours of read speech from children aged 4 to 17. It includes 137,597 utterances from 3,175 speakers: 78,993 from female speakers, 53,927 from male speakers, and 4,677 from speakers with unknown gender information. The audio files for speakers with unknown gender information were removed from the training and test dataset as this would prohibit calculating the gender biases.

This dataset was pre-split into a train, test and validation set. However, this corpus was far larger than the other two corpora and hence to ensure a more equal comparison with the other datasets the size of the training set was reduced from roughly 127 hours to roughly 10 hours. This new dataset was approximately the same size as the other datasets, whilst maintaining the original age and gender distributions from the provided test set. When creating this smaller test set, the aim was to still include as many different speakers as possible. Thus, files were manually selected in such a fashion as to guarantee the correct age and gender distributions but also to maximise the number of different speakers, i.e., files with a shorter duration were preferred over those with a longer duration.

3.2. KidsTALC-v1 Corpus

The KidsTALC-v1 corpus [36] features spontaneous speech from monolingual German children, designed for ASR training to aid in speech development research and therapeutic applications. It includes recordings from approximately 300 children, ranging from kindergarten to elementary school. The elicitation contexts span free play, storytelling, conversational discourse, and read texts, aiming to cover a spectrum of spontaneous language.

The dataset was pre-split into training, test, and validation sets. However, the transcriptions for the test split were not published, for this reason a new split had to be created such that the WER can be calculated. The new test split was created from the training set, maintaining the same age and gender distributions as the validation split.

Moreover, for this corpora, some additional preprocessing was needed to segment the audio files. All of the recordings were longer than 30 seconds, which is the maximum input length for the Whisper model, so they had to be split up into multiple parts. This segmentation was possible using the provided timestamps.

3.3. JASMIN-CGN

The JASMIN-CGN project [37] collected Dutch speech from children, non-natives with various mother tongues, and elderly people. It also included speech from human-machine interactions. From this corpus, only data from groups one and two was selected, which only include native child and teenager (children aged 12 and above) speech. As for the Icelandic dataset, all speakers with unknown age or gender information were discarded.

For the Jasmin set the data was divided into a Dutch and Flemish portion each of which contained a further split into read-speech and human-machine interaction speech. However, each portion had to be further divided into a test, training and validation set. To create the validation and test splits, the same methodology as the KidsTALC dataset was applied, ensuring a balanced representation of male and female speakers, with 2 male and 2 female speakers selected per age group for test and validation sets. The remaining data was used as the training set.

Again this corpora contained audio files that were longer than 30 seconds which had to be segmented, the individual segments were taken from the GitHub [38].

3.4. Overview Corpora

An overview of the datasets in terms of their size and age and gender distributions is provided in Table 1. It is noteworthy that teenage speech, which is defined as children aged 12 or older, was only used for testing purposes, i.e., to see how training impacts the performance of the model on speakers from other age groups. This decision was made based on comparisons of word error rates across different age groups. Adult speech is recognised the best, with teenage speech achieving similar results [21]. In contrast, elderly speech performs worse, and child speech is recognised the least accurately. It is interesting to note that Whisper was trained on 13344 hours of German speech, 16 hours of Icelandic speech and 2077 hours of Dutch speech [22].

Table 1: *Data Summary of Corpora*

Split	Length	Gender		Age			
		#M	#F	3-5	6-8	9-11	12+
IS Train	10h0m	407	407	8	222	561	0
IS Val	1h50m	312	308	0	92	258	270
IS Test	1h50m	315	310	0	94	261	270
DE Train	7h47m	11	11	16	2	4	0
DE Val	1h50m	4	4	4	2	2	0
DE Test	1h48m	4	4	4	2	2	0
NL Train	4h51m	19	25	0	11	33	0
NL Val	1h18m	6	6	0	4	4	4
NL Test	1h19m	6	6	0	4	4	4
VL Train	2h33m	12	15	0	10	17	0
VL Val	1h23m	6	6	0	4	4	4
VL Test	1h19m	6	6	0	4	4	4

4. Experiments

This section outlines the experimental procedures used to evaluate the performance of the Whisper model on child speech recognition. It begins with an overview of the zero-shot testing phase, discussed in Subsection 4.1, where baseline performance is established across the different demographic groups. Subse-

quently, in Subsection 4.2, the finetuning process is explored, where the Whisper model is adapted using transfer learning on child-specific datasets. This section includes details on parameter settings and training methodologies. Finally, in Subsection 4.3, the evaluation phase is discussed, where the metrics of the finetuned model across various age and gender groups are reassessed to quantify improvements.

4.1. Zero-shot Testing

Zero-shot testing, as employed in this experiment, evaluates the performance of the Whisper models on datasets without prior finetuning for child speech recognition. This approach assesses the models' ability to generalise to new languages and age groups without specific adaptation, providing insights into their baseline capabilities and informing subsequent finetuning strategies.

For the initial benchmarking of each dataset (Icelandic, German, Dutch, and Flemish), the WERs were calculated by running Whisper on the test sets. To establish a solid baseline, this was conducted for the following Whisper models: tiny, base, small, medium, large_v1, large_v2, and large_v3. The WER was not only calculated for the corpus as a whole but for individual demographic groups (where applicable for the corresponding dataset): Female (ages 3-5), Female (ages 6-8), Female (ages 9-11), Female (ages 12+), Male (ages 3-5), Male (ages 6-8), Male (ages 9-11), and Male (ages 12+). Then based on the calculated WERs per demographic the gender biases (male vs female) and the age biases (children aged 3-5 vs children aged 6-8 vs children aged 9-11 vs children aged 12+) were calculated. The corresponding benchmarking results can be found in Subsection 5.1.

4.2. Finetuning

After having established the baselines, the next step was then to finetune the models using transfer learning on the child-specific datasets. This process involves using pre-existing knowledge from the Whisper model and adapting it to better accommodate the unique acoustic and linguistic characteristics present in child speech.

This process was conducted for each of the four datasets individually for the corresponding best-performing model from the zero-shot testing. As previously stated, in Subsection 2.3, the low-rank adaptation approach will be taken here. For this method the following parameters were used: r= 32, lora_alpha = 64, lora_dropout=0.05 and, bias="none". Regarding the training phase, the parameters utilised were as follows: batch size of 32 per device, 1 gradient accumulation step, a learning rate of 10^{-4}, 50 warm-up steps, and 10 training epochs with evaluation performed at each epoch. These parameter settings were taken from a paper that also employed LoRA for finetuning Whisper for child speech [39]. Moreover, the temperature parameter was set to 0.0 in Whisper's generation process to have a deterministic generation i.e. only the most probable token is chosen at each step [40].

To prevent overtraining, the training process employed a technique known as early stopping [41]. This approach involved monitoring the validation error after each epoch. The model underwent training for 10 epochs, with the process halted as soon as the validation error surpassed that of the previous epoch.

4.3. Evaluation

Following the finetuning process, the recognition accuracy of the Whisper model was again assessed across various ages and genders. This evaluation aimed to quantify the improvements achieved after adapting the model to child-specific datasets. By analysing recognition accuracy across different demographic groups, including age and gender, the effectiveness of the finetuning process in mitigating biases and improving overall performance can be assessed. The corresponding results can be found in Subsection 5.2.

5. Results & Discussion

In this chapter, the results, i.e. the WERs and biases, are presented. In Subsection 5.1 the initial baseline results after zero-shot testing are presented and discussed. In Subsection 5.2 the results after the finetuning process are presented and compared to the initial results.

5.1. Pre-Finetuning Results

In this subsection, the outcomes of zero-shot testing are discussed. In Subsubsection 5.1.1, the best-performing model based on WER and bias is identified. Subsubsection 5.1.2 compares WER across languages and evaluates the influence of training data volume. Subsubsection 5.1.3 analyses age and gender biases observed in the initial zero-shot results.

5.1.1. Model Selection

Table 2 provides an overview of the zero-shot results for the different model sizes for each language. The best results for each language are highlighted, with the large model size demonstrating the lowest WER for Icelandic (54.91%), Dutch (24.44%), and the overall average (40.44%). For German and Flemish, the large-v3 model boasts the best performance with a WER of 54.11% and 27.60% respectively. These findings highlight the importance of model size in mitigating WER across diverse linguistic contexts. Moreover, when comparing these results to the WERs achieved by Whisper on adult speech Librispeech (see Subsection 2.1), it becomes obvious that these results are in a different order of magnitude, showing that Whisper recognises child speech far worse.[3]

Table 2: *Average WER for Different Model Sizes (%)*

Model Size	IS	DE	NL	VL	Overall Average
Tiny	129.52	98.72	81.51	81.49	97.81
Base	115.05	105.38	72.27	79.74	93.11
Small	96.97	78.24	47.50	53.28	69.00
Medium	80.14	63.28	35.34	39.30	54.51
Large-v1	**54.91**	54.77	**24.44**	27.62	**40.44**
Large-v2	63.90	77.94	32.81	35.99	52.66
Large-v3	54.97	**54.11**	25.63	**27.60**	40.58

Based on the WERs the age and gender biases were then computed as explained in Subsection 2.2. In Table 3 the averaged

[3] Although it should be noted that Whisper was trained significantly more on English speech.

age and gender biases are presented across different model sizes for each language. The best results for each language are emphasised, with the large-v3 model showing the lowest average bias German (12.30%), Dutch (6.29%), Flemish (7.10%), and the overall average (7.84%). The large-v2 model achieved the lowest bias for the Icelandic dataset (5.38%).

Table 3: *Average Bias for Different Model Sizes (%)*

Model Size	IS	DE	NL	VL	Overall Average
Tiny	3.42	25.15	6.18	8.61	10.84
Base	1.58	33.70	8.60	10.96	13.71
Small	4.95	24.63	8.51	12.26	12.59
Medium	7.47	19.38	7.43	10.32	11.15
Large-v1	5.68	12.84	6.52	7.13	8.04
Large-v2	**5.38**	19.86	7.32	11.06	10.91
Large-v3	5.68	**12.53**	**6.29**	**7.10**	**7.84**

There is a notable performance disparity among Whisper models, specifically between large-v2 and both large-v1 and large-v3. Large-v3 benefits from additional training data and architectural improvements, leveraging one million hours of weakly labelled audio and 4 million hours of pseudo-labelled data derived from large-v2, which enhances its performance [32]. The extended training of large-v2 for an additional 2.5 epochs may lead to over-fitting on healthy adult speech, reducing its effectiveness on child speech. Moreover, manual inspection of transcripts revealed that the model transcriptions contained fabricated phrases or sentences not present in the original audio, a phenomenon known as hallucinations [42]. The large-v2 model exhibited a significantly higher incidence of hallucinations compared to the large-v1 and large-v3 models. For instance, in one German transcription, the model generated the phrase ”Ich bin sehr sehr sehr...” repeated 133 times, significantly increasing the WER due to these inaccuracies.

Based on these observations and the results presented in Tables 2 and 3, the large-v3 model was selected for finetuning, as it achieved the best results for bias mitigation and was a close second in recognition performance. This model is also arguably the most state-of-the-art, being the most recently released and trained on the most extensive dataset among the Whisper models. For the remainder of this report, all presented results are for the large-v3 model.

5.1.2. WER Results

There is a substantial disparity between the volumes in languages included in the training data for Whisper’s large-v3 model. Out of the wide variety of languages the model was trained on it contains 13,344 hours of German speech, 16 hours of Icelandic speech, and 2,077 hours of Dutch speech [22]. Based on this it was hypothesised that the model would exhibit the best performance on the German corpus, followed by the Dutch and Flemish corpora, and the poorest performance on the Icelandic corpus. The model’s actual performance, measured in WER is depicted in Figure 1.

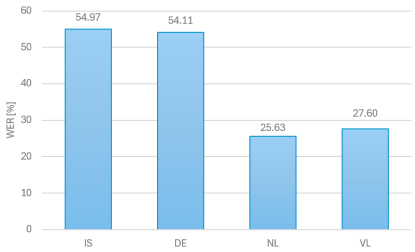


Figure 1: *WER by Language*

The Icelandic corpus yielded the highest WER at 54.97%, which aligns with expectations due to the minimal training data (only 16 hours). This limited exposure to Icelandic speech is insufficient for the model to generalise well, leading to poorer performance. These results are notably worse than those reported by C. Mena et al. [35], ranging between 24.47% and 43.71%. It should be noted, however, that this paper used the Kaldi model alongside a language model and a lexicon.

Contrary to the initial hypothesis, the German corpus did not perform the best despite having the largest amount of training data. The WER for German was 54.11%, only marginally better than Icelandic. Several factors contributed to this unexpected outcome:

- Spontaneous Speech: The German test set exclusively comprised spontaneous speech, which is inherently more variable and complex than read speech. Spontaneous speech includes more disfluencies, colloquialisms, and varying speech patterns, challenging the model that is predominantly trained on read speech. In contrast, the Icelandic dataset contained only read speech, and the Dutch and Flemish datasets included a mix of read and spontaneous speech.
- Child Speech: The German test set uniquely included speech from children aged 3-5, which is more challenging for recognition models due to early developmental speech patterns. Other test sets only included children aged 6-12, whose speech is less variable.
- Over-Training: There are indications of over-training on the German data. This is evidenced by peculiar model behaviours, such as consistently misinterpreting the sound “oh” as “Untertitlung des ZDFs, 2020,” reflecting an over-specialisation to the training data, particularly subtitles from the German public television broadcaster ZDF.[4] This over-specialisation limits the model’s ability to generalise beyond the specific patterns present in the training dataset.

When comparing the results to those of T.B. Patel et al. [43], the large-v3 model outperforms the results obtained using the ESPnet toolkit. Even after applying Speed Perturbation (SP) and Spectral Augmentation (SpecAug), the ESPnet model only achieved a WER of 67.20%.

The large-v3 model performed comparably well on Dutch and Flemish corpora, with WERs of 25.63% and 27.60%, respectively. Flemish, being a dialect of Dutch, shares significant linguistic similarities with Dutch, leading to the model’s similar performance on these two corpora. This finding corroborates with a previous study by Feng et al. [16], which also reported similar WERs for Dutch and Flemish, with Flemish also performing slightly worse.

[4]This misinterpretation occurred 25 times for the large-v3 model

5.1.3. Bias Results

The WER rates were not only calculated for a language as a whole, they were also calculated for each specific age & gender demographic, these results are presented in Appendix A. Based on these WERs the age and gender biases were then calculated. These are presented in Tables 4 and 5.

Table 4: *Bias by Gender [%]*

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

Based on the results presented in Table 4, it is evident that Whisper does not exhibit a consistent bias towards a single gender across different languages. The Icelandic and Flemish datasets show a preference towards female speakers, while the German and Dutch datasets tend to favour male speakers. The Icelandic dataset has a gender bias of 3.75%, however, this difference was deemed not significant enough to conclude the presence of a gender bias that would substantially impact the everyday usability of these tools. On the other hand, the German, Dutch, and Flemish datasets exhibit slightly more pronounced biases, with biases of 6.27%, 6.52%, and 10.30%, respectively. These variations suggest that the model's performance may be influenced by specific demographic and linguistic characteristics inherent to each dataset.

Table 5: *Bias by Age [%]*

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

In Table 5 the age biases are presented.[5] The general trend shows that the younger age groups (3-5 and 6-8) tend to exhibit the highest biases across all languages. The middle age groups (9-11) showed either less bias than the younger age group or no bias. The oldest age group (12+) showed the least bias across all languages. Notably the age biases are much more pronounced than the gender biases, with the German dataset showing the most pronounced biases. The Icelandic dataset is least biased which may be down to the fact that it contains by far the most speakers (625 vs about 8-12 in the other data sets) and hence has the most variability.

5.2. Post-Finetuning Results

In this subsection, the outcomes after applying the LoRA adaption approach that was outlined in Subsection 4.2 are discussed. In Subsubsection 5.2.1, the WER results post-finetuning are presented and compared with pre-finetuning results. Subsubsection 5.2.2 analyses the changes in age and gender biases following the finetuning process. For conciseness, the results in this section are grouped solely by demographic, as it aligns with the primary focus of this research. Detailed results grouped by language are provided in Appendix A.

[5]Note that N/A indicates that this age group does not exist in the dataset, and a bias of 0.00% indicates the lowest WER achieved within this dimension (see Subsection 2.2).

5.2.1. WER Results

The WER results after finetuning show a significant improvement across all demographic groups, indicating the effectiveness of the finetuning process. The relative percentage differences in WER by demographic group are summarised in Table 6 and the pre- and post-finetuning WERs are also visually represented in Figure 2 to give an indication of the absolute values of the achieved WERs.

Table 6: *Relative Percentage Difference in WER by Demographic Group*

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%

From Table 6, several interesting observations can be made. On average, WER improved by nearly 15% across all demographic groups, indicating that the finetuning process was not only successful but also generalised well without over-fitting. The largest improvement in WER is observed for the 'Female' and 'Ages 6-8' demographic groups, both with a reduction of approximately 19%. The 'Ages 12+' group shows the least improvement, with a reduction of only 8.44%. However, this was to be expected as the training set only contained child speech in the age range of 3-11. The main aim of this research was to finetune ASR systems for 'child' speech, and children aged 12 and older were only included in the test set to evaluate how well the post-finetuning model generalises to teenagers.

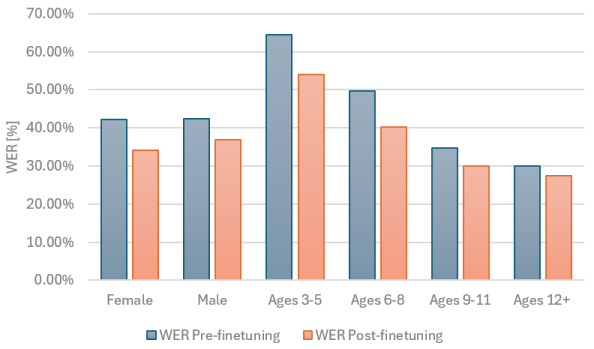


Figure 2: *Comparison of WER Before and After Finetuning Across Demographic Groups*

However, as illustrated in Figure 2, the absolute values of WER remain high, especially when compared to the WERs achieved by Whisper on healthy adult speech. This suggests that further research and refinement in this field are necessary to achieve more competitive WERs for child speech.

5.2.2. *Bias Results*

The bias results following finetuning indicate varied changes across demographic groups. The relative percentage differences in bias by demographic group are summarised in Table 7, with visual representations provided in Figure 3.

Table 7: *Relative Percentage Difference in Bias by Demographic Group*

Demographic Group	Change in Bias
Female	-71.38%
Male	+5.84%
Ages 3-5	-60.80%
Ages 6-8	-42.56%
Ages 9-11	-75.51%
Ages 12+	+68.79%
AVG	-29.27%

From Table 7, it is evident that the average bias reduction across all demographic groups is approximately 29.27%, indicating a generally positive outcome. Notably, excluding the age group '12+' — where an increase in bias was expected due to the deliberate exclusion of any recordings from this demographic — the results show significant bias reductions across other age groups. This suggests effective finetuning for younger age ranges. The most substantial reduction in bias is observed in the 'Ages 9-11' group, with a decrease of 75.51%, and among females, with a decrease of 71.38%. On average, gender biases decreased by 32.77%, while age biases also improved significantly, showing an average decrease of 27.52%. These findings highlight that training on a balanced dataset led to more equitable recognition outcomes.

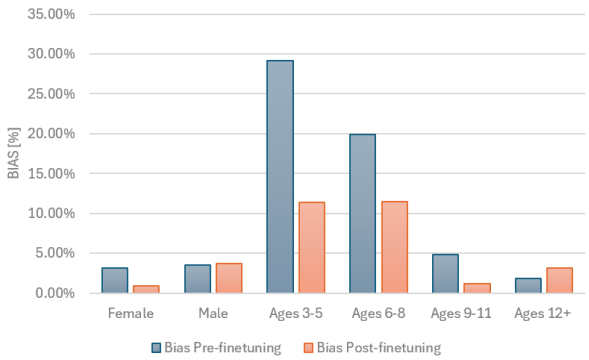


Figure 3: *Comparison of Bias Before and After Finetuning Across Demographic Groups*

From Figure 3, one small anomaly becomes evident: the bias increased very slightly for the male group by 2.84%. Further inspection revealed that this increase was due to a bias flip in the German dataset. Before finetuning, the performance was better for males, whereas after finetuning, it shifted to be better for females. However, since overall bias decreased, the performance now shows more similarity between genders compared to before finetuning.

6. Conclusions and Future Work

Finally in this section some concluding remarks are made in Subsection 6.1 and some ideas for future research areas in the field of Child ASR are proposed in Subsection 6.2.

6.1. Conclusion

This study evaluated the performance of the Whisper ASR model on child speech across multiple languages and demographic groups, focusing on improving WERs and mitigating biases. In the course of this paper the following questions were addressed:

How effectively does the pre-trained Whisper model recognise child speech across different age groups and genders?

The initial zero-shot testing highlighted significant challenges in Whisper model accuracy across different sizes. Specifically, the large-v3 model emerged as the most effective baseline, achieving the lowest average WER of 40.58% across Icelandic, German, Dutch, and Flemish datasets. However, disparities persisted, particularly in recognising spontaneous child speech and addressing age biases, notably in younger age groups (3-5 years).

What age and gender biases exist in the pre-trained Whisper model's recognition of child speech?

Analysis revealed notable biases in the pre-trained Whisper model. Gender biases varied across datasets, with the German and Dutch datasets favouring female speakers and the Icelandic and Flemish datasets favouring male speakers. The average gender bias was 3.36%. The age biases were far more pronounced pronounced with the average bias being 13.97%. These biases are especially pronounced in the younger age groups, with children aged 3-5 having a bias of 29.18%

What changes occur in recognition performance after finetuning the Whisper model with child speech data?

After applying Low-Rank Adaptation (LoRA) finetuning, substantial improvements were observed in recognition performance. The large-v3 model demonstrated enhanced adaptability to spontaneous speech, resulting in an average relative reduction in WER of 15.23% across all datasets. The most impressive results were achieved for the Dutch dataset for which a relative reduction in of 33.39% was achieved, with the post-finetuning WER dropping to 17.69%. Whilst the WERs achieved a still not comparable to those achieved by Whisper on healthy adult speech, the initial results achieved through finetuning with child-specific data are promising.

How do age and gender biases in the Whisper model's recognition of child speech evolve following finetuning?

Finetuning also led to reductions in age and gender biases within the Whisper model. Gender biases decreased by 32.77% on average across datasets, indicating a more balanced recognition of male and female speakers. Age biases also showed improvements, where biases decreased by 27.52% after finetuning. Especially impressive were the reductions in age bias for the youngest age group, where the relative reduction in WER was 60.80%.

These findings have important implications for advancing ASR technology in child-centric applications such as educational

tools and assistive technologies. By improving recognition accuracy and reducing biases, ASR systems like Whisper can better support children's learning and communication needs. This study employs a direct bias mitigation strategy through finetuning on balanced datasets, emphasising the critical role of diverse dataset composition in mitigating bias in speech recognition technology. These insights aim to promote more inclusive practices in future dataset curation and model development, ensuring ASR systems effectively serve diverse user demographics.

6.2. Future Work

The future work in Automatic Speech Recognition (ASR) for child speech entails several key areas of focus, these have been broken down into two parts: recommendations for the current research and additional research directions.

Based on the current research it becomes obvious that there is a need to focus on expanding dataset diversity by creating public speech corpora that feature varied child speech data across different languages, dialects, and age groups to ensure balanced representation. This specific project selected four Germanic languages, it would be interesting to see if the same trends hold true for non-Germanic languages. It would also be interesting to make more extensive use of the Jasmin dataset: one could investigate differences in the performance for HMI vs read speech or investigate the bias in native vs non-native speakers and the bias between regional accents. Moreover, experimenting with different hyperparameters and finetuning techniques would help further improve ASR model adaptation for child speech. On this topic one could also examine the difference in performance between LoRA and other finetuning methods from the PEFT library such as Low-Rank Hadamard Product (LoHa), Low-Rank Kronecker Product (LoKr), and Adaptive Low-Rank Adaptation (AdaLoRA). Additionally, there is a need to develop a more robust method that allows us to assess model biases considering both age and gender. Another key issue that arose was hallucinations and these need to be addressed, which requires refining training data and validation mechanisms. Lastly, one could investigate whether using other state-of-the-art ASR models such as Kaldi or Wav2Vec or other versions of Whisper such as distil-Whisper or faster-Whisper leads to better results.

There also some additional research directions that one could pursue such as investigating and addressing other biases related to dialects, regional accents, non-native accents, ethnicity, race, socioeconomic status, speech rate, and culture. In addition, it would be interesting to perform an intersectional WER analysis which would allow us to identify compounding biases. Moreover, one could explore some advanced bias mitigation techniques, such as adversarial training or data augmentation with synthetic data representing under-represented groups. One could also conduct a longitudinal study that analyses children's speech development over time and then use these insights to create an ASR model that adapts to evolving speech patterns.

7. Responsible Research

When conducting scientific research, researchers inherently assume a significant level of responsibility, necessitating honesty, thoroughness, transparency, independence, and accountability. In order to ensure the ethical and reproducible nature of the research conducted for this Bachelor thesis, the TU Delft Code of Conduct [44] was adhered to. In this section of the report, some ethical considerations are addressed in Subsection 7.1, the reproducability of results is scrutinised in Subsection 7.2, and an overview of the usage of generative AI in this research is provided in Subsection 7.3.

7.1. Ethical Considerations

When undertaking research involving ASR technology, especially with child speech, there are several ethical considerations that must be addressed to ensure the research is conducted responsibly.

While data collection did not fall within the scope of this specific project, it was still ensured that the selected datasets were obtained through ethical means. Notably, attention was paid that stringent efforts were made to secure consent, which is especially crucial for individuals under the age of 18, in adherence to General Data Protection Regulation (GDPR) guidelines [45]. For instance, in situations where parental consent forms were misplaced, as in the case of some speakers from the JASMIN dataset, data pertaining to these individuals was conscientiously discarded[6]. Moreover, the presence of trained speech language therapists or speech language therapy students during the KidsTALC and Jasmin study ensured the ethical conduct of the studies involving child participants. Furthermore for all datasets any personally identifiable information within the speech corpora had been replaced with alphanumeric identifiers.

After discussion with the EEMCS Data Stewards the decision was made to only execute and store data and code on the secure DelftBlue servers. By opting for internal university servers over alternative platforms such as Kaggle or Google Cloud this not only guaranteed the protection of sensitive child speech data but also ensured that the speech corpora, some of which are not publicly available, were not leaked.

Beyond the scope of this project, it is important to contemplate the long-term implications of ASR technology development for children. While the primary aim of this project was to promote inclusivity, it is crucial to continuously evaluate the performance of ASR systems. Vigilance is essential to prevent the inadvertent perpetuation of biases, especially concerning those with atypical speech patterns.

7.2. Reproducibility of Methods

To ensure that the conducted research is reproducible great attention was paid to give a very detailed methodology description. In an attempt to allow others to understand and replicate this methodology the data preprocessing, model adaptation, and evaluation procedures were documented in detail, with reference to specific hyperparameters were appropriate. Throughout the project, the research made use of open-source software and frameworks such as the Whisper model and PyTorch for machine learning tasks, so other researchers can use the same tools to reproduce or build upon the work. Moreover, attention

[6] In the case of certain speakers from the JASMIN dataset, their data was deemed unsuitable for inclusion due to the loss of consent forms by a teacher, under "mysterious" circumstances, despite purported provision by parents. As a result, their data was responsibly discarded.

was paid to using standardised metrics such as Word Error Rate when evaluating the performance such that other researchers can compare results directly. The code used for data processing, training and evaluation of models has been compiled into a GitLab repository.[7]

7.3. Usage of Generative AI Models

During the writing of the Bachelor thesis, three generative AI tools were utilised: ChatGPT, QuillBot, and GitHub Copilot. ChatGPT served primarily for proofreading sections of the report and offering feedback. It also facilitated the conversion of Excel tables to LaTeX format, although this often required significant input to ensure proper formatting. Prompts such as "Please give feedback on my abstract" were provided alongside the abstract, referencing grading rubrics from FeedbackFruits. Quillbot was employed to correct spelling and grammar errors throughout the report, and ensuring a suitable tone.

For coding tasks, GitHub Copilot was extensively used, especially during dataset preprocessing. The approach involved initially mapping out the program flow and breaking it down into separate parts. After documenting these different parts and providing these instructions to Copilot through comments the "generate" prompt was used which often performed astonishingly well and certainly increased the efficiency of writing code.

[7] GitLab Repository: `https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Zhang_Yue/fzeisler-Exploring-state-of-the-art-speech-recognisers-for`

8. References

[1] D. Yu and L. Deng, *Automatic speech recognition*, vol. 1. Springer, 2016.

[2] V. Bhardwaj *et al.*, "Automatic speech recognition (asr) systems for children: A systematic literature review," *Applied Sciences*, vol. 12, p. 4419, Apr. 2022.

[3] M. Gerosa *et al.*, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, pp. 847–860, 10 2007.

[4] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 603–616, 2003.

[5] R. Sobti *et al.*, "Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges," *Multimedia Tools and Applications*, Mar. 2024.

[6] S. S. Gray *et al.*, "Child automatic speech recognition for US English: child interaction with living-room-electronic-devices," in *Proc. 4th Workshop on Child Computer Interaction (WOCCI 2014)*, pp. 21–26, 2014.

[7] R. Southwell *et al.*, "Automatic speech recognition tuned for child speech in the classroom," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Seoul, Korea, Republic of), pp. 12291–12295, IEEE, Apr. 2024.

[8] P. Sheng *et al.*, "Gans for children: A generative data augmentation strategy for children speech recognition," in *INTERSPEECH 2019*, pp. 129–135, 12 2019.

[9] J. Fainberg *et al.*, "Improving children's speech recognition through out-of-domain data augmentation," in *INTERSPEECH 2016*, pp. 1598–1602, 09 2016.

[10] Y. Qian *et al.*, "Improving dnn-based automatic recognition of non-native children speech with adult speech," tech. rep., INTERSPEECH, 09 2016.

[11] R. Tong *et al.*, "Transfer learning for children's speech recognition," in *INTERSPEECH 2017*, pp. 36–39, 12 2017.

[12] M. Matassoni *et al.*, "Non-native children speech recognition through transfer learning," in *INTERSPEECH 2018*, pp. 6229–6233, 04 2018.

[13] R. Jain *et al.*, "Adaptation of whisper models to child speech recognition," in *INTERSPEECH 2023*, pp. 5242–5246, ISCA, Aug. 2023.

[14] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, pp. 349–352 vol. 1, 1996.

[15] M. Abushariah and M. Sawalha, "The effects of speakers gender, age, and region on overall performance of arabic automatic speech recognition systems using the phonetically rich and balanced modern standard arabic speech corpus," 01 2013.

[16] S. Feng *et al.*, "Quantifying bias in automatic speech recognition," 2021.

[17] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (D. Hovy *et al.*, eds.), (Valencia, Spain), pp. 53–59, Association for Computational Linguistics, Apr. 2017.

[18] A. Koenecke *et al.*, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[19] S. Goldwater *et al.*, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, pp. 181–200, 03 2010.

[20] R. Tatman and C. Kasten, "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions," in *Proc. Interspeech 2017*, pp. 934–938, 2017.

[21] S. Feng *et al.*, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, p. 101567, Mar. 2024.

[22] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," Tech. Rep. arXiv:2212.04356, OpenAI, Dec. 2022.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210, IEEE, 2015.

[24] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," 2015.

[25] M. Ravanelli *et al.*, "The pytorch-kaldi speech recognition toolkit," 2019.

[26] M. Ravanelli *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[27] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[28] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," 2020.

[29] Picovoice, "Github - picovoice/speech-to-text-benchmark: speech to text benchmark framework." `https://github.com/Picovoice/speech-to-text-benchmark`, 2018.

[30] Vásquez-Correa *et al.*, "Novel speech recognition systems applied to forensics within child exploitation: Wav2vec2. 0 vs. whisper," *Sensors*, vol. 23, no. 4, p. 1843, 2023.

[31] openai, "Announcing the large-v2 model openai/whisper discussion #661." `https://github.com/openai/whisper/discussions/661`, Dec 2022.

[32] openai, "'large-v3' release openai/whisper discussion #1762." `https://github.com/openai/whisper/discussions/1762`, Nov 2023.

[33] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[34] "Peft (parameter-efficient fine-tuning)." `https://huggingface.co/docs/peft/en/index`, 2024.

[35] H. Mena *et al.*, "Samromur children icelandic speech 1.0." Web Download, 2022.

[36] L. Rumberg *et al.*, "kidsTALC: A Corpus of 3- to 11-year-old German Children's Connected Natural Speech," in *Proc. Interspeech 2022*, pp. 5160–5164, 2022.

[37] 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 Fifth International Conference on Language Resources and Evaluation (LREC'06)* (N. Calzolari *et al.*, eds.), (Genoa, Italy), European Language Resources Association (ELRA), May 2006.

[38] syfengcuhk, "Github - quantifying bias in automatic speech recognition." `https://github.com/syfengcuhk/jasmin`, 2021.

[39] R. Southwell *et al.*, "Automatic speech recognition tuned for child speech in the classroom," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12291–12295, IEEE, 2024.

[40] "Model documentation - whisper." `https://huggingface.co/docs/transformers/en/model_doc/whisper`, 2014.

[41] E. Romero *et al.*, "Feature selection forcing overtraining may help to improve performance," in *Proceedings of the International Joint Conference on Neural Networks, 2003.*, vol. 3, pp. 2181–2186, IEEE, 2003.

[42] A. Koenecke *et al.*, "Careless whisper: Speech-to-text hallucination harms," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, (New York, NY, USA), p. 1672–1681, Association for Computing Machinery, 2024.

[43] T. Patel and O. Scharenborg, "Improving end-to-end models for children's speech recognition," *Applied Sciences*, vol. 14, p. 2353, Mar. 2024.

[44] S. Roeser and S. Copeland, "Tu delft code of conduct: Why what who how," report, Delft University of Technology, 2020.

[45] E. Commission, D.-G. for Justice, and Consumers, *The GDPR – New opportunities, new obligations – What every business needs to know about the EU's General Data Protection Regulation*. Publications Office, 2018.

A. Appendix - Raw Data

A.1. Early Stopping

Table 8 illustrates the behaviours of validation losses over ten epochs for each language. Values in bold indicate the epoch at which the training was halted, as validation loss increased beyond these points.

Table 8: *Validation Losses across Different Epochs*

Epoch	IS	DE	VL	NL
1	0.094	0.056	**0.047**	0.018
2	0.075	0.051	0.123	0.016
3	0.070	0.050	0.024	0.016
4	0.066	**0.049**	0.023	**0.015**
5	0.064	0.050	0.024	0.016
6	0.063	0.051	0.024	0.016
7	**0.062**	0.051	0.024	0.016
8	0.063	0.052	0.024	0.017
9	0.062	0.053	0.024	0.017
10	0.063	0.053	0.025	0.017

A.2. Complete Pre-finetuning Results

This section presents an analysis of zero-shot WERs and bias across different architectures and speaker groups for the four datasets: IS, DE, and NL. The tables detail WER values for each of the seven Whisper models ranging from tiny to large-v3, evaluating performance for both female and male speakers, and across various age groups. Additionally, the bias per group is quantified to assess disparities in model performance.

Table 9: *IS - Zero-shot WER Values across Different Architectures and Speaker Groups (%)*

IS - WER	Tiny	Base	Small	Medium	Large	Large-v2	Large-v3
Female (3-5)	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Female (6-8)	125.15	112.87	89.25	78.98	59.79	66.67	60.05
Female (9-11)	136.12	116.71	103.28	79.95	53.76	63.01	53.76
Female (12+)	125.43	113.22	92.07	69.49	46.34	56.44	45.97
Male (3-5)	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Male (6-8)	138.61	120.66	107.72	94.05	62.39	73.77	62.75
Male (9-11)	129.50	114.55	102.41	87.30	59.43	67.66	59.55
Male (12+)	122.32	112.31	87.11	71.06	47.74	55.82	47.72

Table 10: *IS - Zero-shot WER per group (%)*

IS - WER per group	Tiny	Base	Small	Medium	Large	Large-v2	Large-v3
Female	128.53	114.17	94.76	74.70	50.53	60.66	51.41
Male	128.58	115.06	97.02	81.80	55.04	63.97	55.16
Ages 6-8	132.28	117.00	99.04	86.96	61.17	70.44	61.48
Ages 9-11	132.64	115.58	102.82	83.81	56.74	65.45	56.80
Ages 12+	123.93	112.78	89.68	70.25	47.02	56.14	46.82
AVG	129.19	114.92	96.66	79.50	54.10	63.33	54.33

Table 11: *IS - Bias per group (%)*

IS - Bias per group	Tiny	Base	Small	Medium	Large	Large-v2	Large-v3
Female	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Male	0.05	0.89	2.26	7.10	4.51	3.31	3.75
Ages 6-8	8.35	4.22	9.36	16.71	14.15	14.30	14.66
Ages 9-11	8.71	2.80	13.14	13.56	9.72	9.31	9.98
Ages 12+	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AVG	3.42	1.58	4.95	7.47	5.68	5.38	5.68

Table 12: *DE - Zero-shot WER Values across Different Architectures and Speaker Groups (%)*

DE - WER	Tiny	Base	Small	Medium	Large	Large-v2	Large-v3
Female (3-5)	116.12	123.17	95.76	75.18	67.82	98.60	67.33
Female (6-8)	133.92	152.77	91.29	77.99	66.18	97.75	65.49
Female (9-11)	70.73	59.57	41.99	30.64	37.59	47.78	36.43
Female (12+)	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Male (3-5)	108.72	124.60	90.49	80.90	59.07	90.50	57.50
Male (6-8)	112.83	130.02	117.17	80.79	64.86	80.19	64.82
Male (9-11)	50.00	42.16	32.72	34.19	33.09	52.82	33.09
Male (12+)	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 13: *DE - Zero-shot WER per group (%)*

DE - WER per group	Tiny	Base	Small	Medium	Large	Large-v2	Large-v3
Female	118.43	127.86	90.21	72.39	64.75	94.51	64.47
Male	105.55	120.12	95.96	76.91	59.07	83.75	58.20
Ages 3-5	113.70	123.62	94.04	77.15	64.81	96.06	64.37
Ages 6-8	126.11	144.38	100.91	79.00	65.70	91.50	65.25
Ages 9-11	63.47	53.61	38.77	31.89	36.00	49.52	35.19
AVG	105.45	113.92	83.98	67.47	58.07	83.07	57.50

Table 14: *DE - Bias per group (%)*

DE - Bias per group	Tiny	Base	Small	Medium	Large	Large-v2	Large-v3
Female	12.88	7.74	0.00	0.00	5.68	10.76	6.27
Male	0.00	0.00	5.75	4.52	0.00	0.00	0.00
Ages 3-5	50.23	70.01	55.27	45.26	28.81	46.54	29.18
Ages 6-8	62.64	90.77	62.14	47.11	29.70	41.98	30.06
Ages 9-11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AVG	**25.15**	**33.70**	**24.63**	**19.38**	**12.84**	**19.86**	**13.10**

Table 15: *NL - Zero-shot WER Values across Different Architectures and Speaker Groups (%)*

NL - WER	Tiny	Base	Small	Medium	Large	Large-v2	Large-v3
Female (3-5)	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Female (6-8)	83.84	87.56	60.54	49.34	35.13	42.48	34.16
Female (9-11)	90.83	77.21	50.11	38.08	33.91	38.80	33.91
Female (12+)	78.54	65.85	39.73	29.00	19.72	26.78	20.11
Male (3-5)	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Male (6-8)	90.25	83.41	57.86	44.63	21.07	42.00	32.59
Male (9-11)	78.01	63.61	46.93	29.74	21.07	28.25	21.25
Male (12+)	67.60	55.95	29.81	21.25	15.74	18.52	15.74

Table 16: *NL - Zero-shot WER per group (%)*

NL - WER per group	Tiny	Base	Small	Medium	Large	Large-v2	Large-v3
Female	85.19	77.62	50.81	39.41	30.50	36.80	30.28
Male	79.57	68.50	46.31	32.72	23.80	30.51	23.76
Ages 6-8	87.16	85.42	59.16	46.90	33.98	42.23	33.35
Ages 9-11	84.32	70.30	48.49	33.85	27.39	33.44	27.48
Ages 12+	73.09	60.92	34.79	25.14	17.74	22.67	17.94
AVG	81.87	72.55	47.91	35.60	26.68	33.13	26.56

Table 17: *NL - Bias per group (%)*

NL - Bias per group	Tiny	Base	Small	Medium	Large	Large-v2	Large-v3
Female	5.62	9.12	4.50	6.69	6.70	6.29	6.52
Male	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ages 6-8	14.07	24.50	24.37	21.76	16.24	19.56	15.41
Ages 9-11	11.23	9.38	13.70	8.71	9.65	10.77	9.54
Ages 12+	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AVG	6.18	8.60	8.51	7.43	6.52	7.32	6.29

Table 18: *VL - Zero-shot WER Values across Different Architectures and Speaker Groups (%)*

VL - WER	Tiny	Base	Small	Medium	Large	Large-v2	Large-v3
Female (3-5)	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Female (6-8)	81.43	109.63	64.01	48.50	32.45	32.49	32.45
Female (9-11)	68.31	57.70	35.20	24.41	17.67	22.68	17.67
Female (12+)	75.56	64.13	40.41	26.27	17.18	25.87	17.18
Male (3-5)	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Male (6-8)	94.05	102.69	78.34	66.99	45.14	69.22	45.12
Male (9-11)	74.34	67.73	39.38	27.95	20.74	25.05	20.74
Male (12+)	95.23	76.55	62.34	41.67	32.56	40.63	32.42

Table 19: *VL - Zero-shot WER per group (%)*

VL - WER per group	Tiny	Base	Small	Medium	Large	Large-v2	Large-v3
Female	75.44	77.73	46.92	33.28	22.51	29.23	22.51
Male	88.19	82.21	60.21	45.50	32.87	44.93	32.81
Ages 6-8	87.83	106.11	71.28	57.88	38.89	54.08	38.88
Ages 9-11	71.57	63.13	37.46	26.32	19.33	23.96	19.33
Ages 12+	85.62	70.48	51.63	34.15	25.05	33.42	24.97
AVG	81.73	79.93	53.50	39.43	27.73	37.12	27.70

Table 20: *VL - Bias per group (%)*

VL - Bias per group	Tiny	Base	Small	Medium	Large	Large-v2	Large-v3
Female	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Male	12.75	4.48	13.29	12.22	10.36	15.70	10.30
Ages 6-8	16.26	42.98	33.82	31.56	19.56	30.12	19.55
Ages 9-11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ages 12+	14.05	7.35	14.17	7.83	5.72	9.46	5.64
AVG	8.61	10.96	12.26	10.32	7.13	11.06	7.10

A.3. Overview large-v3 Pre-finetuning Results

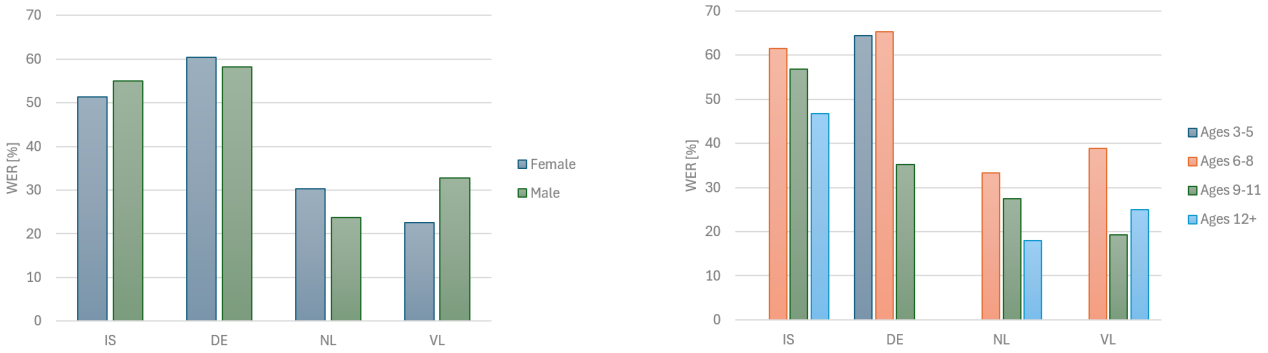
This section provides a more in-depth overview of the pre-finetuning performance of the Large-v3 model, which was chosen for finetuning. It presents the WERs and associated biases, segmented by gender and age across the four datasets (IS, DE, NL, and VL). Additionally, the results for the WERs are displayed graphically in Figure 4.

Table 21: *WER and Bias for Large-v3 by Gender [%]*

Gender	IS		DE		NL		VL	
	WER	Bias	WER	Bias	WER	Bias	WER	Bias
Female	51.41	0.00	64.47	6.27	30.28	6.52	22.51	0.00
Male	55.16	3.75	58.20	0.00	23.76	0.00	32.81	10.30

Table 22: *WER and Bias for Large-v3 by Age [%]*

Age	IS		DE		NL		VL	
	WER	Bias	WER	Bias	WER	Bias	WER	Bias
3-5	N/A	N/A	64.37	29.18	N/A	N/A	N/A	N/A
6-8	61.48	14.66	65.25	30.06	33.35	15.41	38.88	19.55
9-11	56.80	9.98	35.19	0.00	27.48	9.54	19.33	0.00
12+	46.82	0.00	N/A	N/A	17.94	0.00	24.97	5.64



(a) *WER by Gender* (b) *WER by Age*

Figure 4: *WER Analysis*

A.4. Comparison Pre-finetuning vs Post-finetuning Results

This section compares the pre-finetuning and post-finetuning performance of the large-v3 model. It presents the WERs and associated biases for various groups, segmented by gender and age across the four datasets (IS, DE, NL, and VL). The results are displayed in tabular form to allow for a quick comparison of before (Table 23) and after (Table 24).

Table 23: *Pre-finetuning WER per Group [%] and Bias per Group [%]*

WER	IS	DE	NL	VL	AVG
Female	51.41%	64.47%	30.28%	22.51%	**42.17%**
Male	55.16%	58.20%	23.76%	32.81%	**42.48%**
Ages 3-5	N/A	64.37%	N/A	N/A	**64.37%**
Ages 6-8	61.48%	65.25%	33.35%	38.88%	**49.74%**
Ages 9-11	56.80%	35.19%	27.48%	19.33%	**34.70%**
Ages 12+	46.82%	N/A	17.94%	24.97%	**29.91%**
AVG	**54.33%**	**57.50%**	**26.56%**	**27.70%**	-

Bias	IS	DE	NL	VL	AVG
Female	0.00%	6.27%	6.52%	0.00%	**3.20%**
Male	3.75%	0.00%	0.00%	10.30%	**3.51%**
Ages 3-5	N/A	29.18%	N/A	N/A	**29.18%**
Ages 6-8	14.66%	30.06%	15.41%	19.55%	**19.92%**
Ages 9-11	9.98%	0.00%	9.54%	0.00%	**4.88%**
Ages 12+	0.00%	N/A	0.00%	5.64%	**1.88%**
AVG	**5.68%**	**13.10%**	**6.29%**	**7.10%**	-

Table 24: *Post-finetuning WER per Group [%] and Bias per Group [%]*

WER	IS	DE	NL	VL	AVG
Female	40.67%	52.69%	20.04%	22.83%	**34.06%**
Male	43.36%	55.19%	16.38%	32.51%	**36.86%**
Ages 3-5	N/A	53.97%	N/A	N/A	**53.97%**
Ages 6-8	44.95%	55.11%	23.17%	37.71%	**40.24%**
Ages 9-11	44.01%	42.53%	15.55%	17.86%	**29.99%**
Ages 12+	39.23%	N/A	15.67%	27.26%	**27.39%**
AVG	**42.44%**	**51.90%**	**17.69%**	**28.84%**	-

Bias	IS	DE	NL	VL	AVG
Female	0.00%	0.00%	3.66%	0.00%	**0.92%**
Male	2.69%	2.50%	0.00%	9.68%	**3.72%**
Ages 3-5	N/A	11.44%	N/A	N/A	**11.44%**
Ages 6-8	5.72%	12.58%	7.62%	19.85%	**11.44%**
Ages 9-11	4.78%	0.00%	0.00%	0.00%	**1.20%**
Ages 12+	0.00%	N/A	0.12%	9.40%	**3.17%**
AVG	**2.64%**	**5.30%**	**2.28%**	**7.79%**	-

Table 25: *Relative Percentage Difference in WER and Bias by Language [%]*

Language	Change in WER
IS	-21.88%
DE	-9.74%
NL	-33.39%
VL	+4.10%
AVG	-15.23%

Language	Change in Bias
IS	-53.54%
DE	-59.52%
NL	-63.78%
VL	+9.69%
AVG	-37.87%

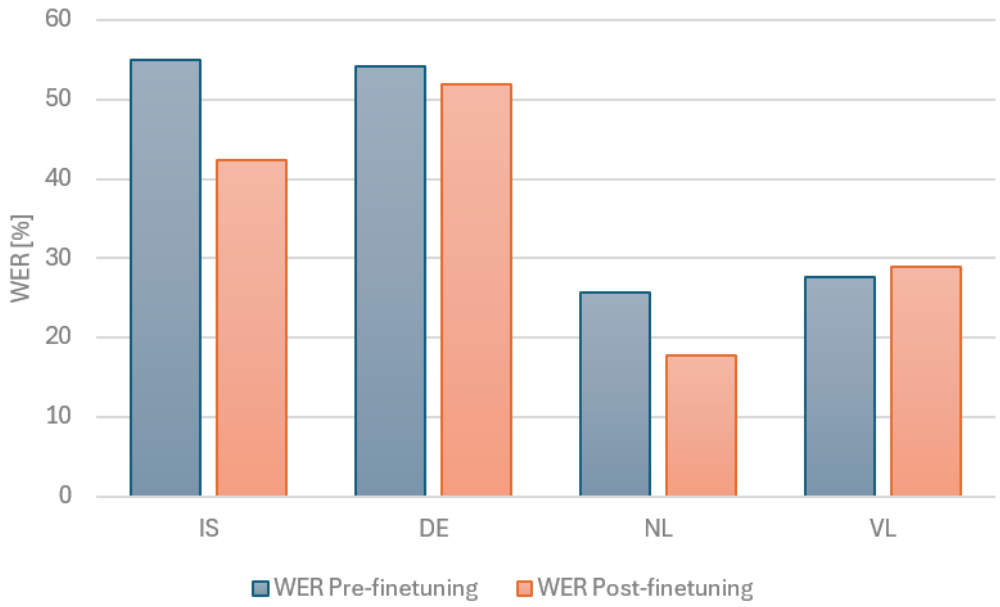


Figure 5: *Effect on WER of finetuning, sorted by language*