From KidWhisper to SplitWhisper: Investigating Demographic Fine-Tuning in ASR for Children

Arnout Van Elk[1], *Timon Crouzen*[1], *Lola Vissers*[1]

[1]RU s1054707, [1]RU s1034392, [1]RU s1052874

arnout.vanelk@ru.nl, timon.crouzen@ru.nl, lola.vissers@ru.nl

Abstract

It is becoming more common that children's automatic speech recognition (ASR) systems are being used and thus that research has increased. But it is far from the level of adult ASR systems. It is generally not well known how age and gender influence model performance, especially on state of the art models like KidWhisper. In this study we investigate how age and gender influence model performance, how specific age and gender finetuning affects performance, and if age classification is feasible on children's speech. Our results support that age has some influence over the Word Error Rate (WER) for children's ASR as the WER dropped from 29.08 WER for younger children to 12.45 WER for older children. While our results also showcase that specialised fine-tuning based on age and gender only give minimal ASR performance improvement. We also show that age classifiers are feasible on children's speech.

Index Terms: ASR, Whisper, KidWhisper, Finetuning, Child Speech Recognition

1. Introduction

Automatic Speech Recognition (ASR) is the process by which speech is transformed into text. A lot of improvements have been introduced and achieved over the past few years. However, this research mainly focused on the development of ASR systems for adult speech, the recognition of children's speech has received less attention [1], [2]. This is due to many reasons. For instance, collecting speech data from children is more challenging in comparison to collecting speech data from adults, mostly because of privacy concerns. Furthermore, ASR for children is more challenging due to the difference in children's speech and adult speech [2], [3]. Differences include acoustic features, frequency, pitch, and linguistic features [2], [3]. Additionally there is a lot of speech variability among children, especially at younger ages [2]. Consequently, the field of ASR for children remains relatively unexplored, with several research opportunities still available to close the gap between adult ASR models and children's ASR models [1], [2], [4].

A recent advancement in improving ASR for children is KidWhisper, introduced by A. Attia et al [5]. The KidWhisper approach enhances children's ASR by effectively preprocessing child speech data and fine-tuning the Whisper language model on this data. Whisper [6] is mainly trained on adult speech data and performs well on adult speech data.

In this study, we begin by replicating the methodology presented in the original KidWhisper paper [5]. We extend this research by looking into speech variations among children based on age and gender. Specifically, we aim to further improve child speech recognition by implementing gender- and age-specific models. In addition, we also focus on implementing gender and age classifiers to make the pipeline of gender- and age-specific models complete. In this paper we plan to address three different research questions:

1. How does the performance of the KidWhisper model vary across different age and gender groups of children?
2. To what extent does specialized fine-tuning on age and gender improve performance for child speech?
3. Can children's age and gender be reliably identified using a classifier trained on speech data?

2. Related Work

While advancements like KidWhisper have pushed the boundaries of ASR for children, some argue that the physiological and cognitive aspects of children's learning process are fundamental. As such we aim to find what effect physiological aspects such as gender and age have on the performance of an ASR Model, which is dependent on the available children speech corpora. In this paper we will talk about two such children speech corpora. The first of which is the My Science Tutor project corpus (MyST). Pradhan et al. [7] reported a Word Error Rate (WER) of 11.6 when testing on the MyST test set. They used an end-to-end transformer model, fine-tuned using the MyST training set and trained on the LibriSpeech model using utterances less than 30 seconds.

A. Attia et al. [5], which serves as inspiration for this study, report a WER of 33.85 and 28.47 on the test set of the CSLU Kids Scripted and Spontaneous respectively. They achieved this by using a small English only Whisper model trained on the MyST corpus. However, when the authors trained on a combination of MyST and CSLU, they achieved a WER of 2.59 (Scripted) and 27.16 (Spontaneous).

Lastly we look at the results that Fan et al. [1] report with their benchmarking of different childrens ASR models. For the Whisper-small model they reported a WER of 1.8 on the CSLU OGI Corpus, and a WER of 1.5 for the Whisper-medium model.

Different papers report results of models where they considered such physiological and cognitive aspects while other papers do not consider these aspects. Singh et al. [4] report a WER of 54.0 on the CSLU kids dataset with a self-supervised Wav2Vec model before fine-tuning and a 32.4 WER after fine-tuning. They also reported the results of a weakly supervised Whisper model with a WER of 26.28 before fine-tuning and 25.2 after fine-tuning. Singh et al. [4] further indicate that these SFMs show a non-negligible sensitivity to age as an physiological factor. In comparison they show that the SFMs have a lesser sensitivity to gender.

Shivakumar et al. [3] also reported the results of different

models pertaining to the age of the children who's speech the models were tested on. However, Shivakumar et al. [3] trained their models on adult speech before testing them on the OGI Kids Corpus. Additionally, the models they used wrre End-to-End learning models and Neural networks. The architectures they used are for example, Factorized Time Delay Neural Network (TDNN-F) HMM Systems, Residual Neural Networks (ResNet) and Time-Depth Seperable (TDS) Convolution Networks and Transformers. For their TDNN-F DNN-HMM model they found a WER of 53.55 without the MySt finetuning and a 30.40 WER with finetuning. And for their best performing model the TDS + CTC + 4-gram language model they found a WER of 37.32 without finetuning and 33.54 with. For the kindergarten speech they speciy that the TDNN-F DNN-HMM Model shows an 17.31% relative improvement in WER when finetuned on MyST, while the TDS + CTC + 4-gram Language Model has an improvement of 4.08 on kindergarten level.

3. Methodology

3.1. Datasets

To investigate the impact of age and gender on ASR for children, we considered several datasets. Our initial intention was to use the MyST dataset [8], one of the largest available datasets of child speech. MyST has mainly been used in the original KidWhisper study, making it an ideal candidate for a reproduction and extension of the study. However, when we inspected the metadata, we found that MyST lacks the information about age and gender of the children. Unfortunately, this information is necessary for answering our research questions; therefore, the MyST dataset is unsuitable for this study.

After inspecting the MyST dataset, we looked into using the CSLU Kids Speech corpus [9]. This dataset is smaller than MyST. However, it still consists of a relatively large and diverse sample of child speech, over 1,100 children, grades 0 to 10, around 100 hours of recorded audio. Fortunately, this dataset did include metadata such as gender and age. It did not specifically include the age of the children, but it did include the grade level. The gender labels were retrieved from [10]. In this study we decided to treat the grade as an age indicator. Specifically, we focus on grades 0 through 5, which correspond approximately to children aged 4 to 10 years. This range has been chosen because the gender labels were only available for this grade range [10]. The original KidWhisper paper used grades 3–7.

In this study, we decided to treat the grade as an age indicator, which introduces some imprecision, as children within the same grade can vary in age by up to a year or more. This limitation is discussed further in the discussion.

The CSLU dataset includes speech from children in educational settings. We followed preprocessing procedures aligned with those in the KidWhisper study, including filtering for audio quality and excluding utterances with incomplete metadata. Due to Whisper's pipeline requirements, the Spontaneous speech can only be used for testing and not for fine-tuning, as the audio files are too long. Therefore, throughout the whole paper, only the Scripted speech will be used for fine-tuning. Moreover, because of the different task context of the Spontaneous dataset, it provides a solid way to test the generalization of a model to different circumstances.

3.2. Models

3.2.1. Whisper

We used the Whisper models developed by OpenAI [6] to perform our experiments. Specifically, we chose the small variant of Whisper, which is in alignment with the original KidWhisper study. The Whisper models are transformer architectures pretrained on a large and diverse corpus of adult speech data. This broad pre-training allows Whisper to generalize across different domains. This is a property that KidWhisper uses to adjust the model on child speech data.

3.2.2. Age & Gender Classsifier

We developed a custom Residual Neural Network with two heads which can be enabled to predict binary age labels, binary gender labels, or both. A class-weighted Binary Cross-Entropy loss function was used for each head. When both heads were enabled, the two losses were combined, each weighted by their exponential moving average. We trained and validated the model on only the Scripted CSLU set.

3.3. Experimental Procedure

The experimental procedure consists of multiple steps to replicate and extend the KidWhisper study[1]. All experiments were performed on the Ponyland server, which was selected for its GPU infrastructure and its direct access to CSLU dataset.

The first stage involved forking the GitHub repository associated with the KidWhisper paper. The repository provided implementation details for pre-processing and fine-tuning Whisper models on child speech data. We faced many challenges in integrating and adapting their code to our experimental setup, such as setting up the specific JSON file structure that KidWhisper requires. Another issue is that the CSLU dataset hosted on the Linguistic Data Consortium lacks the transcription files for the Scripted data. This was a challenge as the Scripted data was necessary for finetuning the Whisper models. We solved this by contacting the authors of the KidWhisper paper, they have provided us with the transcriptions.

To evaluate the reproducibility of the KidWhisper paper, we replicated its main experiments using the same Whisper (small, English) model and same pre-processing and fine-tuning steps. However, there are some differences in our pipeline, the main difference is the dataset which has been used. In the original paper the unfiltered CSLU dataset had been used with grades 3-7. As there were only gender labels available for grades 0-5, we decided to use this dataset. This resulted in a smaller fine-tuning dataset with overall younger children, which likely impacted final performance. As we wanted to optimize the limited data we were working with, we filtered the low-quality entries based on the manual labels included in the dataset. This was not done with CSLU in the original KidWhisper paper.

Another challenge was how to divide the age groups. As we were already dealing with a relatively small dataset, we decided to only compare a binary split in the age groups. While the grade to split at is arbitrary (we chose grade 3, the middle grade), it still provides us with a broadly "younger" and broadly "older" group to compare.

After having solved all the challenges we started experimenting with KidWhisper on CSLU data with multiple settings and configurations. With this information, we could evaluate

[1]Code can be found on GitHub: `https://github.com/TimonC/Split-Whisper`

performance variations based on age and gender specialization. The following models have been tested:

1. Baseline Whisper: The Whisper model (without any fine-tuning) was used to establish baseline performance on ASR for children. The Whisper model has been trained on adult datasets and was expected to underperform on children's speech.
2. KidWhisper MyST: This model refers to the publicly available KidWhisper model (small, English) that had already been fine-tuned on the MyST children's speech dataset, as presented in the original KidWhisper paper. We did not reproduce the MyST training ourselves, but used the pre-trained checkpoint released by the authors. This model serves as a baseline for children's speech, already adapted to children's speech through fine-tuning on MyST.
3. KidWhisper MyST + CSLU: To further specialize the KidWhisper model for our evaluation setting, we fine-tuned the KidWhisper MyST model on a subset of the CSLU Kids Speech corpus (grades 0 to 5). This additional fine-tuning step was aimed at improving performance on our specific evaluation splits, for both Scripted and Spontaneous speech settings. The same pre-processing and training steps have been applied as described in the KidWhisper paper.
4. Fine-tuning based on age and gender: The full dataset was split into two age groups, younger children and older children, based on grade metadata. Additionally, it has also been split based on gender, boys and girls. Lastly, there have also been splits based on both age and gender: older boys, older girls, younger boys and younger girls. KidWhisper (+MyST) models were fine-tuned on each subset. The goal was to find out whether age and/or gender specialization leads to an improvement.

In addition to these experiments, a multi-task residual network was trained to predict age and gender labels on the CSLU Scripted dataset. This was based on the idea that specialized KidWhisper models would require classification to match the right transcription model to the data. In line with this idea, the classifier uses the Whisper log-mel audio features. Only the CSLU Scripted dataset was used to keep feature length roughly consistent. Custom masking was applied to remove the significant padding that the Whisper feature extractor applied to the short sound fragments.

Due to the custom masking, and due to the lack of segmentation or attention mechanisms in our classifier, the model was not well suited to training on the Spontaneous dataset. For simplicity we therefore only considered CSLU Scripted. Moreover, we justified this decision with the following assumption: that the different task context in CSLU Spontaneous is not as relevant for age and gender classification as it is with ASR.

During development, it was observed that simply adding the two losses would lead to one task dominating (gender). This is a common challenge in multi-task learning, which is often addressed with dynamic loss weighting strategies ([11], [12]). In line with this, a simple exponential moving average of each loss was used to weight the losses dynamically, so that the larger loss does not dominate the smaller loss during training.

3.4. Metrics

For the evaluation of the models in this study, two metrics were used: *Word Error Rate (WER)* for the ASR tasks, and a class-weighted binary classification accuracy for the classifier. For both models the train, validation and test sets had no overlap in speakers.

3.4.1. WER

The main metric for evaluating the performance of ASR models is the WER, this is a standard metric in speech recognition that measures the similarity between the predicted transcription and the ground truth text. WER is calculated as:

$$\text{WER} = \frac{S + D + I}{N} \tag{1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference transcript.

The WER was computed using normalized text (e.g., removal of punctuation and standardization of casing). Both the Scripted and Spontaneous subsets of the CSLU dataset have been evaluated.

3.4.2. Classification accuracy

For the classification tasks (predicting speaker age group, gender, or both), we considered the binary accuracy on the validation set. The validation accuracy was not used in training, but only to pick the best performing model and report the results. To address class imbalance, the acccuracy was class-weighted:

$$\text{Weighted Accuracy} = \frac{1}{C} \frac{\sum_{c=1}^{C} w_c \cdot \text{Correct}_c}{\sum_{c=1}^{C} w_c \cdot \text{Total}_c} \tag{2}$$

Where w_c is the inverse of the class frequency.

4. Results

This section presents the results of the experiments. Performance is analyzed across different age and gender subgroups, with separate evaluations on the Scripted and Spontaneous subsets of the CSLU dataset. The results will be divided into four parts, the first part reflects on the reproduction of the KidWhisper paper, the other three sections are consistent with the three research questions.

4.1. Reproduction KidWhisper paper

Results are reported in terms of WER on both Scripted and Spontaneous speech CSLU datasets, see table 1.

Without any fine-tuning, the baseline Whisper model achieved a WER of 22.62 on Scripted CSLU and 34.47 on Spontaneous speech. These scores are a bit higher than the paper's reported baselines of 21.31 (Scripted) and 32.00 (Spontaneous). This is likely the result of differences in the subset we consider, most likely due to our filtering of the lower-quality entries.

When using the imported KidWhisper model (i.e., Whisper fine-tuned on MyST), we observed an improvement in Spontaneous speech WER (28.99) compared to the baseline (34.47), aligning with the KidWhisper paper (28.47). Interestingly, however, the Scripted WER in our case was nearly identical to the baseline (22.80 vs. 22.62), whereas the original paper reported a significant degradation (33.85).

One of the most signficant results in the original paper had been the stark improvement on the Spontaneous dataset on for KidWhisper vs Whisper, both for the version trained on MyST

only and the one including CSLU. In table 1 we see a comparable improvement. Moreover, as we can compare the performance between age groups, we can see that this improvement in KidWhisper is primarily for younger children. This indicates that the generalizable features learned by KidWhisper are primarily for children under grade 3, or younger than 6. When finetuning on CSLU, the older group performs better.

4.2. How does the performance of the KidWhisper model vary across different age and gender groups of children?

To investigate how the KidWhisper model performs across different age and gender groups, we analyzed word error rates (WERs) on both Scripted and Spontaneous CSLU speech for children segmented by age (younger vs. older) and gender (girls vs. boys). We compared three model variants: (1) Whisper without any fine-tuning, (2) KidWhisper fine-tuned on MyST, and (3) KidWhisper further fine-tuned on the CSLU corpus (grades 0–5). The results can be found in table 1, table 2 and table 3.

4.2.1. Age

For the baseline Whisper model, recognition performance was significantly better for older children than for younger ones. For Scripted CSLU speech, WER dropped from 29.06 (younger) to 12.45 (older), while for Spontaneous speech it dropped from 37.76 to 27.86. This suggests that the model (trained primarily on adult speech) struggles more with speech patterns of younger children.

Fine-tuning on MyST improves performance for younger children in Spontaneous settings (37.76 to 28.26), but had limited effect on Scripted data (29.06 to 28.15). For older children, the Scripted WER increased slightly (12.45 to 14.29) while Spontaneous performance slightly worsened (27.86 to 30.46), likely due to domain mismatch between MyST and CSLU.

Further fine-tuning on CSLU significantly improved results across both age groups. For younger children, the WER on Scripted speech dropped to 8.73, while older children achieved 3.00. On Spontaneous speech, the improvements were more modest, with WERs of 31.93 (younger) and 25.19 (older). For the Scripted WER. the result indicate overfitting, as was also the case in the original KidWhisper paper. However, as the Spontaneous dataset provides a better test case of gauging how well models the learn generalizible child-specific features, the improvements here are more significant.

	Scripted			Spontaneous		
Model	All	Younger	Older	All	Younger	Older
Whisper	22.64	29.06	12.45	34.47	37.76	27.86
KidWhisper MyST	22.80	28.15	14.29	**28.99**	**28.26**	30.46
KidWhisper MyST + CSLU	**6.51**	**8.73**	**3.00**	29.69	31.93	**25.19**

Table 1: *Test WER seperated by age groups of the CSLU Kids Speech dataset.*

4.2.2. Gender

Differences in gender were also visible. For the baseline Whisper model, younger girls and boys both had high WERs on Scripted data (34.03 and 33.95, respectively), but older girls performed better than older boys (13.11 vs. 18.23). Spontaneous WERs were high, especially for younger boys (44.28).

With MyST fine-tuning, younger boys saw little change in Spontaneous WER (33.49), while younger girls showed improvement (25.34). For older children, boys benefited more from fine-tuning, for both the scriped and Spontaneous the WER dropped, while girls showed an increase in WER in both cases.

After additional CSLU fine-tuning, WERs for both genders improved substantially. Older girls achieved the lowest Scripted WER of 2.57, closely followed by older boys at 3.48. Younger girls and boys also improved significantly (11.52 and 12.81). On Spontaneous speech, however, results remained relatively high and variable, particularly for younger groups, with girls at 32.09 and boys at 36.96.

Overall, the performance of KidWhisper models varies across both age and gender. Age appears to be the more dominant factor: older children consistently outperform younger ones, likely due to more stable and adult-like speech patterns. In line with this, the generalizable improvements identified in the original Kidwhisper paper appear to primarily be for this more difficult-to-transcribe subsection, i.e. younger children. Gender differences are also present but less consistent, possibly influenced by model biases. These findings suggest that fine-tuning on age and gender data is important for optimizing child speech recognition performance.

	Girl			Boy		
Model	All	Younger	Older	All	Younger	Older
Whisper	21.00	34.03	**13.11**	24.29	33.95	18.23
KidWhisper MyST	22.32	33.29	**15.68**	23.27	33.49	16.86
KidWhisper MyST + CSLU	5.95	11.52	**2.57**	7.08	12.81	3.48

Table 2: *Test WER separated by age and gender groups, for CSLU Scripted data.*

	Girl			Boy		
Model	All	Younger	Older	All	Younger	Older
Whisper	30.28	35.59	**26.01**	38.57	44.28	33.99
KidWhisper MyST	27.65	**25.34**	29.51	30.30	35.94	25.78
KidWhisper MyST + CSLU	28.55	32.09	**25.71**	30.81	36.96	25.87

Table 3: *Test WER separated by age and gender groups, for CSLU Spontaneous data.*

4.3. To what extent does specialized fine-tuning on age and gender improve performance for child speech?

To evaluate the impact of specialized fine-tuning on ASR for children, we analyze WER across models fine-tuned on different age and gender subsets of the CSLU dataset. The base KidWhisper (pretrained on MYST) is further fine-tuned on CSLU Scripted data. We explore whether tailoring models to specific age groups (younger: grades 0–2; older: grades 3–5) and gender (boys, girls) improves recognition performance. The results can be found in table 4, table 5 and table 6.

	Scripted			Spontaneous		
Model	All	Younger	Older	All	Younger	Older
Whisper	22.64	29.06	**12.45**	34.47	37.76	**27.86**
KidWhisper MyST	22.80	28.15	**14.29**	28.99	**28.26**	30.46
KidWhisper MyST + CSLU	6.51	8.73	**3.00**	29.69	31.93	**25.19**
YoungerWhisper	5.76	8.23	**1.84**	26.50	28.30	**22.89**
OlderWhisper	7.62	10.41	**3.19**	26.67	28.17	**23.63**

Table 4: *Test WER seperated by age groups of the CSLU Kids Speech dataset.*

In contrast to expectation, younger-child training data does not always lead to worse performance. For Scripted speech, YoungWhisper outperforms OldWhisper when tested on age groups (8.23 vs. 10.41 WER for younger; 1.84 vs. 3.19 WER for older), and also performs better overall (5.76 vs. 7.62). For Spontaneous speech, YoungWhisper continues to have a slightly better WER, achieving 26.50 WER overall versus OldWhisper's 26.67. These differences are not huge, but they suggest that fine-tuning on data specific to age can be useful.

Specialization based on gender shows some patterns. On Scripted speech, performance differences between BoyWhisper and GirlWhisper are relatively minor, with BoyWhisper

	Girl			Boy		
Model	All	Younger	Older	All	Younger	Older
Whisper	21.00	34.03	**13.11**	24.29	33.95	18.23
KidWhisper MyST	22.32	33.29	**15.68**	23.27	33.49	16.86
KidWhisper MyST + CSLU	5.95	11.52	**2.57**	7.08	12.81	3.48
GirlWhisper	6.08	12.06	**2.45**	6.31	11.76	2.90
BoyWhisper	6.18	11.46	2.98	5.93	11.43	**2.49**
YoungerGirlWhisper	6.10	12.20	**2.41**	7.03	13.67	2.86
OlderGirlWhisper	6.64	13.61	**2.41**	7.23	12.68	3.81
YoungerBoyWhisper	6.61	12.47	3.06	6.44	11.96	**2.98**
OlderBoyWhisper	6.66	11.79	**3.55**	7.38	13.08	3.81

Table 5: *Test WER separated by age and gender groups, for Scripted data.*

achieving slightly lower WERs overall. However, on Spontaneous speech, BoyWhisper consistently outperforms GirlWhisper across all subsets. This suggests that fine-tuning based on gender has more effect on Spontaneous speech.
Looking at models trained on combined age-and-gender subgroups, such as OlderBoyWhisper, YoungerGirlWhisper, we observe no significant gains over the specialized models. For Scripted speech, WERs are very close across all four combinations, indicating minimal difference between younger vs. older or male vs. female. For Spontaneous speech, OlderBoyWhisper performs best overall but only little compared to OlderGirlWhisper, YoungerBoyWhisper, and YoungerGirlWhisper.

In summary, in some cases specialized fine-tuning improves recognition performance over general-purpose child models. However, the expected advantage is not strongly supported.

	Girl			Boy		
Model	All	Younger	Older	All	Younger	Older
Whisper	30.28	35.59	**26.01**	38.57	44.28	33.99
KidWhisper MyST	27.65	**25.34**	29.51	30.30	35.94	25.78
KidWhisper MyST + CSLU	28.55	32.09	**25.71**	30.81	36.96	25.87
GirlWhisper	33.00	34.46	31.83	34.74	38.69	**31.57**
BoyWhisper	29.21	31.55	**27.33**	32.07	36.70	28.36
YoungerGirlWhisper	29.26	32.13	**26.95**	31.81	37.58	27.18
OlderGirlWhisper	34.39	**33.28**	35.28	35.70	37.95	33.90
YoungerBoyWhisper	34.33	35.94	**33.03**	36.92	40.17	34.31
OlderBoyWhisper	26.66	28.33	**25.32**	29.36	33.01	26.44

Table 6: *Test WER separated by age and gender groups, for Spontaneous data.*

4.4. Can children's age and gender be reliably identified using a classifier trained on speech data?

We investigated whether children's age groups can be reliably identified from speech using classifiers trained on Whisper's log-mel audio features. To this end, trained three Residual NN variants on the Whisper log-mel audio features CSLU Scripted dataset: an AgeClassifier for predicting age group (younger vs. older), a GenderClassifier for predicting gender (girl vs. boy), and a combined AgeGenderClassifier trained to predict both simultaneously. All models use class-weighted Binary Cross-Entropy (BCE) loss to handle imbalanced labels, with the AgeGenderClassifier trained using a dynamically weighted sum of both losses to prevent one task from dominating during optimization. The results can be seen in Figure 1.

In terms of accuracy, the AgeGenderClassifier achieves the highest classification performance overall, reaching 79% accuracy on older children. However, its accuracy on younger children drops to 61%, underperforming compared to the AgeClassifier. The AgeClassifier has a more consistent accuracy of 72% across both age groups. This suggests that joint learning may negatively impact the generalization for younger children. One interpretation is that the model, by integrating gender prediction, introduces a bias that distorts the age representation in younger children. In contrast, for older children, gender characteristics tend to be more pronounced, and the model can benefit from the additional gender signal to improve age classification. The GenderClassifier shows modest and stable performance, with accuracy around 65% across both age groups. This lower performance underscores the difficulty of predicting gender from child speech. In summary, age classification from speech is feasible with reasonable accuracy. While joint learning with gender may increases the performance in some subgroups, especially older children, it can also impair performance on others.

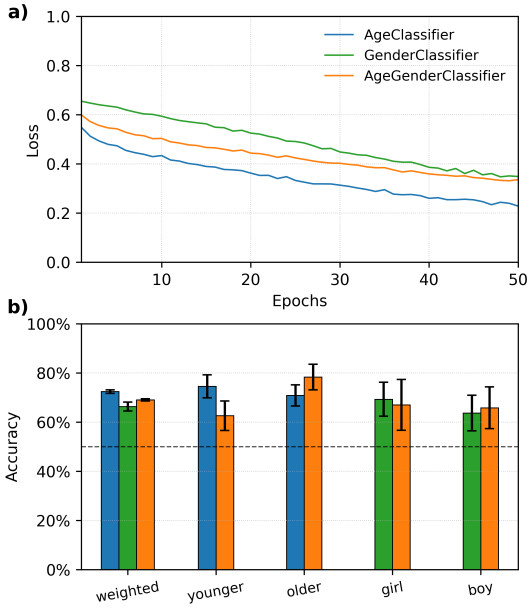

Figure 1: *Training a multi-task residual neural network on the CSLU Scripted dataset.* ***a)*** *BCE loss for one example training run. Results are shown for each task variant of the classifier.* ***b)*** *Classification accuracy on different subsets of the data (weighted includes all validation data) for each model variant, averaged over 5 runs with the standard deviation shown. The model with highest weighted accuracy was used, with performance usually peaking within 20 epochs for all models. The dotted line indicates the baseline chance accuracy.*

5. Discussion

This study explored the performance ASR models on children's speech by building upon the KidWhisper framework. Specifically, we asked: (1) how ASR performance varies across age and gender groups; (2) whether age- and gender-specific fine-tuning improves recognition; and (3) whether age and gender can be predicted from children's speech. Through these research questions, we replicated and extended the existing work in this field.

5.1. How does the performance of the KidWhisper model vary across different age and gender groups of children?

One of the patterns we observed is that age is a significant factor influencing ASR performance. Whisper, pre-trained on

adult speech, performed worse on younger children (e.g., 29.06 WER Scripted vs. 12.45 for older children). This suggests that younger children's speech differs significantly from the adult data Whisper is trained on. This highlights the importance of domain adaptation.

Gender differences, on the other hand, were more subtle. While some performance differences existed, such as older girls slightly outperforming older boys in Scripted tasks, these did not generalize consistently. This could be due to the Whisper model's sensitivity to cues that differ by age more significantly than by gender. These findings indicate that age is a more impactful factor than gender when it comes to fine-tuning ASR for children.

5.2. To what extent does specialized fine-tuning on age and gender improve performance for child speech?

We hypothesized that fine-tuning ASR models on age- and gender-specific data would lead to performance improvements. While specialized fine-tuning did lead to better results in some configurations (e.g., YoungWhisper outperforming OldWhisper), not all results were consistent.

This suggests that while personalization holds promise, the splits (e.g., four age-gender combinations) may have limited data available per subset, which limits the effectiveness of the fine-tuning process. A solution might be to explore strategies where models learn from all data jointly but attend more to age- or gender-specific features during prediction.

In summary, while specialized fine-tuning is promising, it seems to context-dependent. Future research should explore smarter ways of using demographic labels.

5.3. Can children's age and gender be reliably identified using a classifier trained on speech data?

The results suggest that age classification from children's speech is viable. Notably, the model trained only on age achieved more balanced accuracy across age groups, while the joint model performed better for older children but struggled with younger ones. This indicates that incorporating gender information can enhance performance when gender cues are more acoustically distinct but may hinder learning when such cues are less pronounced, as is likely the case with younger children [13], [14].

This interaction implies a trade-off in joint learning: while extra information like gender can distinguish the model's representation space, they could also introduce biases that interfere. The under performance on younger children suggests that the joint model may be overfitting to gender-specific patterns that are not yet reliably present in early speech development[14], or that the model may be encountering feature entanglement. However, the more stable performance of the age-only model points to a more robust representation of age-related features that is less sensitive to these more complex variables.

The instability introduced by multi-task learning is a common issue, requiring careful structuring of the cost functions to overcome [11]. While it is possible that a more refined multi-task cost function weighting could lead to an overall better performance compared to a single-task model, the review in [12] indicates that such strategies don't reliably lead to improvements in performance. Whether we are simply not applying the right strategies, or whether multi-task strategies inherently lead to instability and not to improvements, is a topic that requires more research. For the development of domain-adapted models for children, it is crucial to know if and when we should make use of multi-task learning.

5.4. Reflection

The whole group was quite new to the field of ASR, we learned a lot from this project, it was interesting. First, we were quite surprised by the practical difficulty of working with real-world speech data. From limited metadata to incomplete transcripts and poor documentation in research codebases.

Second, we learned a lot about the domain of child speech. Child speech data is noisy, for example, children sometimes stop talking in de middle of the sentence. It has different features than adult speech data. These features challenge models in ways that adult speech does not.

Lastly, we learned that ASR is not just about building better models, it is about building appropriate models. ASR for children is not just an optimization challenge but an inclusion challenge.

5.5. Limitations and future work

This study faced some important limitations. First, due to metadata constraints, we could only use a subset of the CSLU data (grades 0–5), likely limiting performance compared to the original KidWhisper paper. Moreover, using grade as an indicator for age is noisy. In future work, it would be valuable to use more accurate demographic information.

Additionally, Spontaneous speech recognition remains a bottleneck. Most improvements occurred in Scripted settings, which is understandable given that fine-tuning was done on Scripted data. This calls for future methods that can handle longer, more natural utterances, possibly through segmenting long sequences.

6. Conclusion

This project explored how age and gender affect ASR performance on children's speech, building upon the KidWhisper framework. Our results show that age is a dominant factor influencing recognition accuracy. While gender has a smaller and less consistent effect. Fine-tuning on age and gender subsets resulted some benefits but proved limited by data availability and task complexity. The classification models (age, gender, both) performed adequate. Overall, this work highlights both the promise and challenges of changing ASR systems to better serve children, underscoring the need for inclusive approaches in speech technology.

7. Authors Contribution

The team began by exploring the KidWhisper pipeline. Lola focused on getting the pipeline to run, first with the MyST dataset and later with CSLU. At the same time, Arnout and Timon worked on developing the classification models. Regular check-ins were held to coordinate progress. When the CSLU dataset appeared incomplete, Arnout contacted the original authors to gain full access. Timon then continued improving the code in a new repository and resolved several bugs in the KidWhisper pipeline, and setting up the classifier runs. After experiments were underway, Lola and Arnout worked on writing the report, Arnout focused on the literature review, while Lola handled the remaining sections.

8. References

[1] R. Fan, N. B. Shankar, and A. Alwan, "Benchmarking Children's ASR with Supervised and Self-supervised Speech Foundation Models," 2024.

[2] G. Yeung and A. Alwan, "On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children," in *Interspeech 2018*. ISCA, 2018, pp. 1661–1665.

[3] P. G. Shivakumar and S. Narayanan, "End-to-End Neural Systems for Automatic Children Speech Recognition: An Empirical Study," 2021.

[4] V. P. Singh, M. Sahidullah, and T. Kinnunen, "Causal Analysis of ASR Errors for Children: Quantifying the Impact of Physiological, Cognitive, and Extrinsic Factors," 2025.

[5] A. A. Attia, J. Liu, W. Ai, D. Demszky, and C. Espy-Wilson, "Kid-Whisper: Towards Bridging the Performance Gap in Automatic Speech Recognition for Children VS. Adults," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, no. 1, pp. 74–80, 2024.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 28 492–28 518.

[7] S. S. Pradhan, R. A. Cole, and W. H. Ward, "My Science Tutor (MyST) – A Large Corpus of Children's Conversational Speech," 2023.

[8] W. Ward, S. Pradhan, and R. Cole, "MyST Children's Conversational Speech," 2021.

[9] Shobaki, Khaldoun, Hosom, John-Paul, and Cole, Ronald Allan, "CSLU: Kids' Speech Version 1.1," 2007.

[10] V. P. Singh, M. Sahidullah, and T. Kinnunen, *ChildAugment: Data Augmentation Methods for Zero-Resource Children's Speaker Verification*. the journal of acoustical society of America (JASA) (under review), 2024.

[11] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] T. Gong, T. Lee, C. Stephenson, V. Renduchintala, S. Padhy, A. Ndirango, G. Keskin, and O. H. Elibol, "A comparison of loss weighting strategies for multi task learning in deep neural networks," *IEEE Access*, vol. 7, pp. 141 627–141 632, 2019.

[13] S. P. Whiteside and C. H. and, "Some acoustic characteristics in the voices of 6- to 10-year-old children and adults: a comparative sex and developmental perspective," *Logopedics Phoniatrics Vocology*, vol. 25, no. 3, pp. 122–132, 2000.

[14] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 4, pp. 1011–1021, 2011.