Improving automatic speech recognition for Dutch children with DLD using cross-lingual voice conversion

Author - Jannes Kelso
janneskelso@gmail.com
Supervisor - Yuanyuan Zhang
Professor - Odette Scharenborg

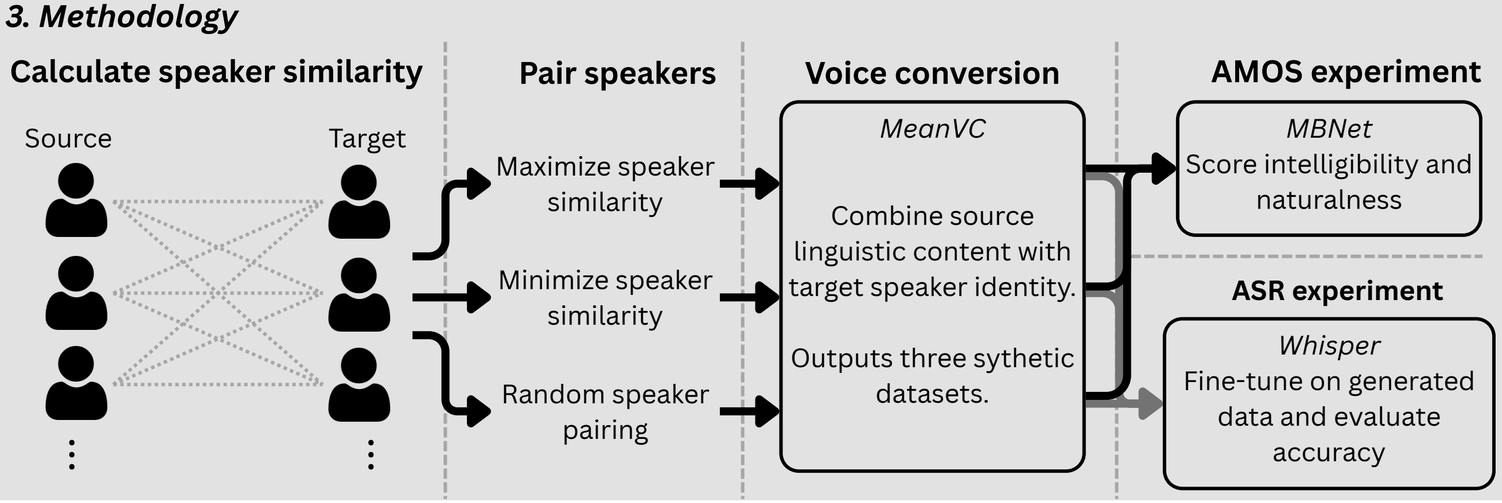
1. Introduction

- Automatic Speech Recognition (ASR) enables children with Developmental Language Disabilities (DLD) to communicate more.
- But ASR models are only as good as their training data, and not much speech data of Dutch children with DLD exists.
- Voice Conversion (VC) preserves the linguistic content of a source utterance, while transferring speaker identity from a target utterance.
- By using using VC to transfer the identity of English children with Speech Sound Disorders (SSD) onto the content of Dutch children, we can generate new data.
- Previous work [1] found VC-generated data can improve *typical* child speech. We hope to improve ASR for DLD child speech by training with this additional data.

2. Research Question

Can **cross-lingual, child-to-child VC** be used to **generate Dutch DLD child speech** as a means of **improving ASR** and what effect does **source-target speaker similarity** have on ASR performance?

3. Methodology



4. Evaluation Methods

ASR Accuracy

- Synthesized data is meant to improve ASR
- So we train an ASR model and test its accuracy against a baseline using WER
- Word Error Rate (WER): compares transcript of input with output for word errors

Automatic Mean Opinion Scores (AMOS)

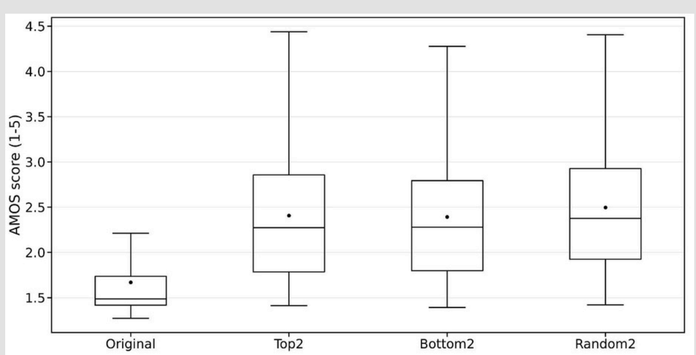
- Score naturalness and intelligibility from 1 to 5
- Trained on human evaluations of speech
- Objective, compared to human MOS
- We use mean bias network (MBNet)

5. Results

Name	WER	Outliers*	WER excluding outliers
Baseline_Model	52.3	2.4	37.4
Original_Model	**34.2**	1.6	28.1
Top_Model	78.3	3.2	34.2
Original_Top_Model	78.2	2.4	33.8
Bottom_Model	79.7	2.4	35.6
Original_Bottom_Model	50.6	2.4	**27.8**
Random_Model	68.8	4.8	31.5
Original_Random_Model	80.9	3.2	35.1

*Percentage of test utterances whose WER exceeded 100%.

WER results for ASR configurations.
All values are percentages



AMOS results for pairing configurations.

6. Conclusion

- Cross-lingual VC from English SSD child speech cannot replace or supplement real Dutch DLD recordings for ASR fine-tuning.
- Low-similarity pairings outperformed high similarity pairings.
- Previous work [1] on *typical* child speech found the opposite.
- DLD child speech is harder to generalize from than typical speech, due in part to its linguistic and phonological diversity.
- Fine-tuning on synthesized data introduced insertion loops.
- If one could filter out bad synthesized data, it may be possible to reduce these errors and use VC to improve ASR for child DLD speech

[1] Yuanyuan Zhang, Zhengjun Yue, Tanvina Patel, and Odette Scharenborg. Improving child speech recognition with augmented child-like speech. In Interspeech 2024, pages 5183–5187, September 2024. arXiv:2406.10284 [cs].