Improving Automatic Speech Recognition for Dutch Children with Developmental Language Disorder Using Cross-Lingual Child-to-Child Voice Conversion

Author: Julie van Montfrans (j.a.vanmontfrans@student.tudelft.nl) , daily supervisor: YuanYuan Zhang, responsible professor: Odette Scharenborg

1. Introduction

Automatic speech recognition (ASR) systems struggle with speech of children with developmental language disorder (DLD) [1].

DLD affects children's language development, making communication and learning more difficult.

Reasons why ASR systems perform poorly on child-with-DLD-speech:
- ASR systems are trained mainly on typical adult speech
- There is a scarcity of speech from Dutch children with DLD

Solution approach: data augmentation through cross-lingual child-to-child voice conversion (VC). VC is a technique that transformas a source's speaker's voice into that of a target speaker while preserving the linguistic content of the original utterance [2].

2. Research Questions

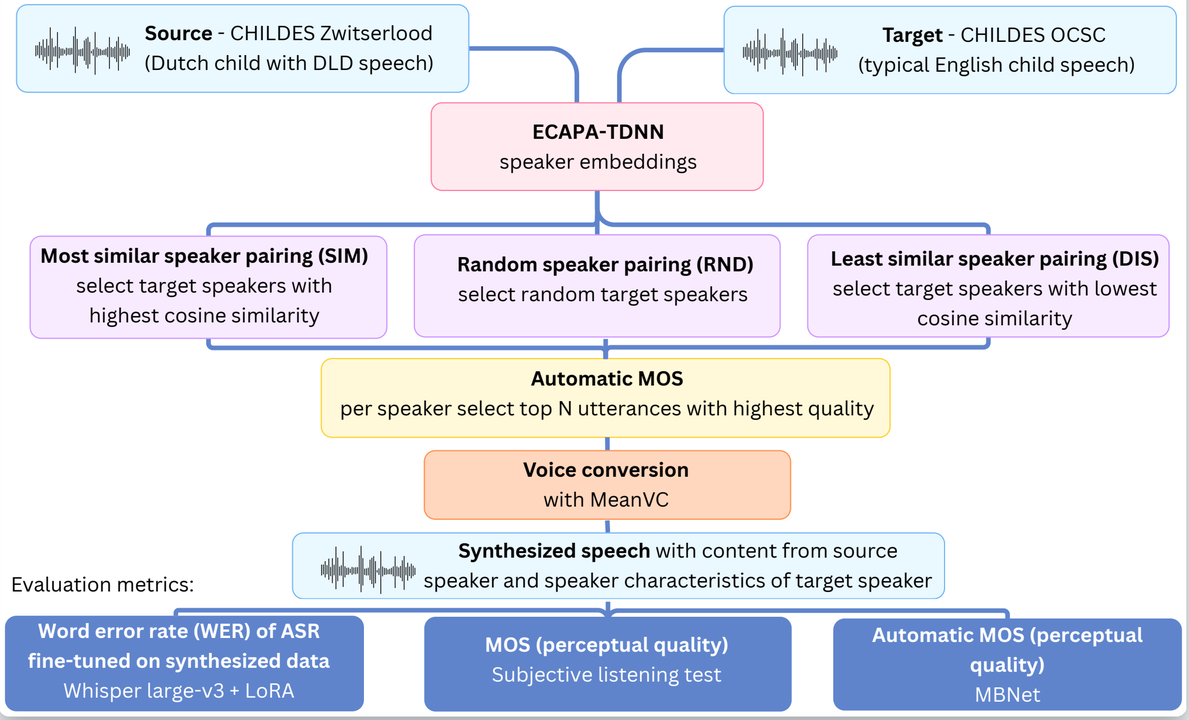
Primary Research Question (RQ):

To what extent can cross-lingual child-to-child VC using non-Dutch children's speech improve ASR performance for Dutch children with DLD?

Sub-questions:

RQ1: To what extent does cross-lingual VC help to improve the performance of ASR for Dutch children with DLD, compared to a baseline consisting of data without augmentation?
RQ2: What is the quality of VC-generated speech, as measured by subjective Mean Opinion Score (MOS)?
RQ3: What is the quality of VC-generated speech, as measured by automatic MOS prediction (AMOS)?
RQ4: To what extent does source-target speaker similarity (similar, random, dissimilar) influence the quality of VC-generated speech?

3. Methodology



4.1 Results - ASR performance

Table 1: WERs per ASR fine-tuning configuration. Bold incidates the lowest WER

Model	Fine-tuning data	Speaker pairing strategy	WER%
ZO	-	-	46.5
ORG	source speech	-	**31.6**
ORG+DIS-2	source + synthetic	Most similar	49.5
ORG+RND-2	source + synthetic	Random	39.1
ORG+SIM-2	source + synthetic	Least similar	33.6

4.2 Results - subjective MOS

Table 2: Mean ratings (1-5) scale for intelligibility, childlikeness, and naturalness across conditions

Condition	Intelligibility	Childlikeness	Naturalness
Source	3.92	4.17	3.9
Most similar	1.93	2.31	1.85
Random	1.92	2.34	1.93
Least similar	1.96	2.34	1.91

4.3 Results - AMOS

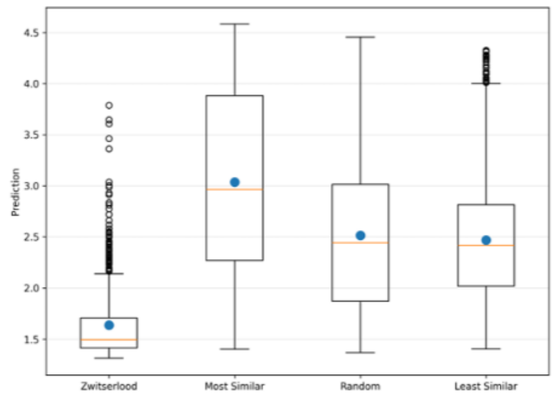


Figure 1: AMOS prediction scores across source-target speaker conditions. The blue filled markers indicate the mean MOS per condition.

5. Conclusion

Cross-lingual child-to-child VC did not improve ASR performance for Dutch children with DLD.

RQ1: The ASR fine-tuned on the synthesized speech did not lead to better performance than an ASR fine-tuned on real speech from Dutch children with DLD.
RQ2: According to MOS, the VC-generated speech was of lower perceptual quality than the source speech.
RQ3: AMOS assigned higher quality scores to the synthetic speech than to the source speech. This contradicts the MOS findings.
RQ4: Dissimilar speaker pairing yielded the best ASR performance. The nfluence of speaker pairing on perceptual quality was relatively limited. Thus the perceptual quality and WER quality are not aligned.

Future work

- Test multiple VC models and ASR architectures for generalizable findings.
- Investigate the effect of using other languages for cross-lingual VC.

References

[1] X. Wan. Improving the performance of automatic speech recognition for children with developmental language disorders. Master's thesis, Delft University of Technology, 2025.
[2] Y. Zhang, Z. Yue, T. Patel, and O. Scharenborg. Improving child speech recognition with augmented child-like speech. In Proceedings of Interspeech 2024, pages 5183–5187, 2024.