Improving ASR for Dutch Children with DLD Through Voice Conversion of Dysarthric Adult Speech

Yulia Waarsenburg

Supervisor: YuanYuan Zhang Professor: Odette Scharenborg

TU Delft Delft University of Technology

Introduction

- ASR performs well for adult speech but **struggles with children**, especially those with DLD.
- Collecting real DLD data is restricted by **privacy regulations, ethics, and annotation issues**.
- Synthetic speech via Voice Conversion (VC) to expand and diversify training data.

Research Questions

- **To what extent does converting Dutch dysarthric adult speech into Dutch child speech via monolingual VC improve ASR performance for Dutch children with DLD?**
  - To what extent are the overlapping acoustic characteristics between Dutch adult dysarthric speech and Dutch child DLD speech preserved during the voice conversion process?
  - Which severity level of adult dysarthria gives the most optimal voice-converted training data for recognizing the speech of Dutch children with DLD?

Acoustic Similarities: Dysarthria & DLD

- **Intelligibility:** Reduced overall clarity of speech.
- **Pronunciation:** Sound substitutions and imprecise consonant articulation.
- **Fluency:** Slower speaking rates accompanied by pauses or hesitations.
- **Prosody:** Atypical rhythm, stress patterns, and intonation.

Methodology

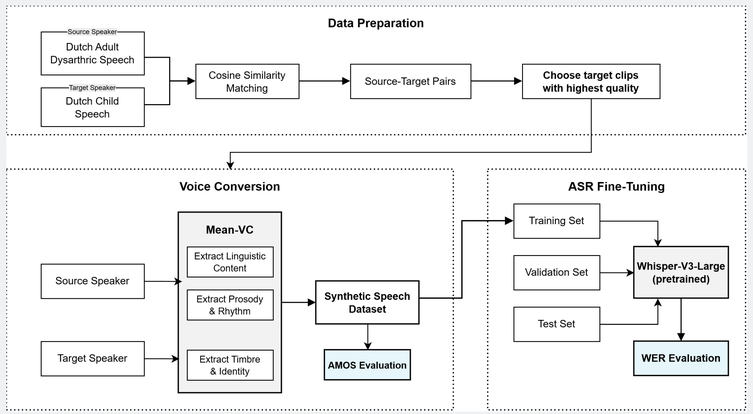


Figure: Overview of the proposed experimental pipeline: data preparation, Mean-VC voice conversion, and ASR fine-tuning.

Datasets

Group	N	M/F	Intel.
Mild	45	29/16	80%–95%
Moderate	23	15/8	50%–79%
Severe	6	4/2	10%–49%
Total	**74**	**48/26**	**10%–100%**

Table 1: COPAS dataset (N = 74).

Split	N	M/F	Age
Train (70%)	42	32/10	8.38
Test (20%)	12	9/3	8.36
Val. (10%)	6	4/2	8.41
Total	**60**	**45/15**	**8.38**

Table 2: Zwitserlood DLD dataset.

Age	Reg.	N	M/F
Child (<13)	NL	77	37/40
Child (<13)	VL	87	43/44
Total	**NL/VL**	**164**	**80/84**

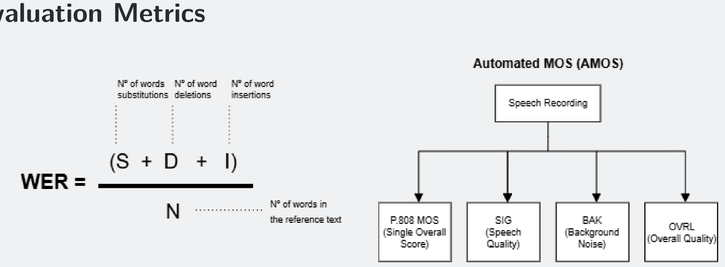
Table 3: JASMIN-CGN dataset.

Speech Type	Duration	Source
Read	3.3 hrs	CGN prompts
Spontaneous	0.4 hrs	–
Total	**3.7 hrs**	**1 spk (severe)**

Table 4: DysOne dataset.

Zwitserlood DLD speech is used for ASR evaluation; COPAS and DysOne provide source dysarthric adult speech for VC; JASMIN provides target speech for VC.

Evaluation Metrics



Results: Speech Quality

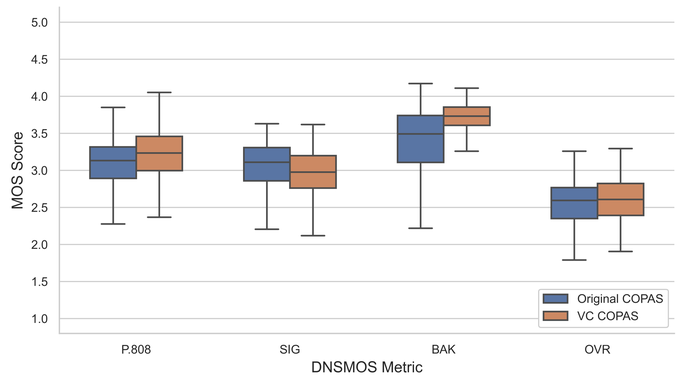


Figure: **Distribution of DNSMOS metrics:** COPAS dataset.

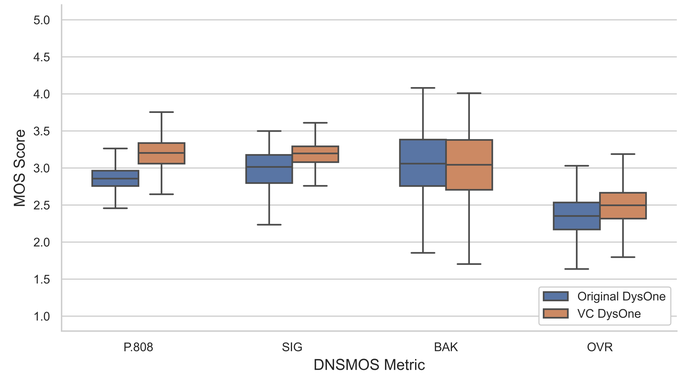


Figure: **Distribution of DNSMOS metrics:** DysOne dataset.

Results: Word Error Rate (WER)

Table: WER (%) of models fine-tuned on VC-generated and DLD data. The best overall performance are highlighted in bold.

Model ID	Description	WER (%)
Baseline		
Base-PT	Pretrained	46.5
Base-DLD	Zwitserlood DLD data	**31.6**
Finetune on VC Speech Only		
VC-Copas-R	VC Copas read speech	**37.4**
VC-Copas-All	VC Copas read and spontaneous speech	42.3
VC-DysOne	VC DysOne speech	41.2
Finetune on DLD + VC Speech		
DLD-VC-Copas-R	VC Copas read + DLD	42.3
DLD-VC-Copas-All	VC Copas read and spontaneous + DLD	38.3
DLD-VC-DysOne	VC DysOne + DLD	57.5
Control Experiments		
Orig-DysOne	Unconverted DysOne	37.0
DLD-Orig-DysOne	Unconverted DysOne + DLD	**33.5**

Conclusions

- While outperforming zero-shot baselines, it yields limited overall improvements to ASR performance.
- VC struggles to preserve essential acoustic characteristics in cases of severe speech impairment.
- Dysarthric speech serves as a viable training alternative when DLD data is unavailable.