Improving Automatic Speech Recognition for Dutch Children with Developmental Language Disorder using Synthetic Data

Synthetic Speech Augmentation Using Text-to-Speech

AUTHORS

Wendy Hu
Responsible Professor: Odette Scharenborg
Supervisor: YuanYuan Zhang

AFFILIATIONS

EEMCS Faculty
Delft University of Technology
June 2026

1. INTRODUCTION & MOTIVATION

Automatic Speech Recognition (ASR) systems now achieve near-human accuracy for typical adult speech, yet remain unreliable for children with Developmental Language Disorder (DLD) — a neurodevelopmental condition affecting grammar, phonology, and morphology.

Core problem:

Collecting real DLD child speech is severely constrained by GDPR, ethical protections, and high annotation cost.

Research question:

To what extent does synthetic Dutch child speech generated from DLD-inspired text improve ASR for children with DLD?

Two Experiment Factors:

- Phoneme Error Probability
- Speaker variability (1, 5, 10)

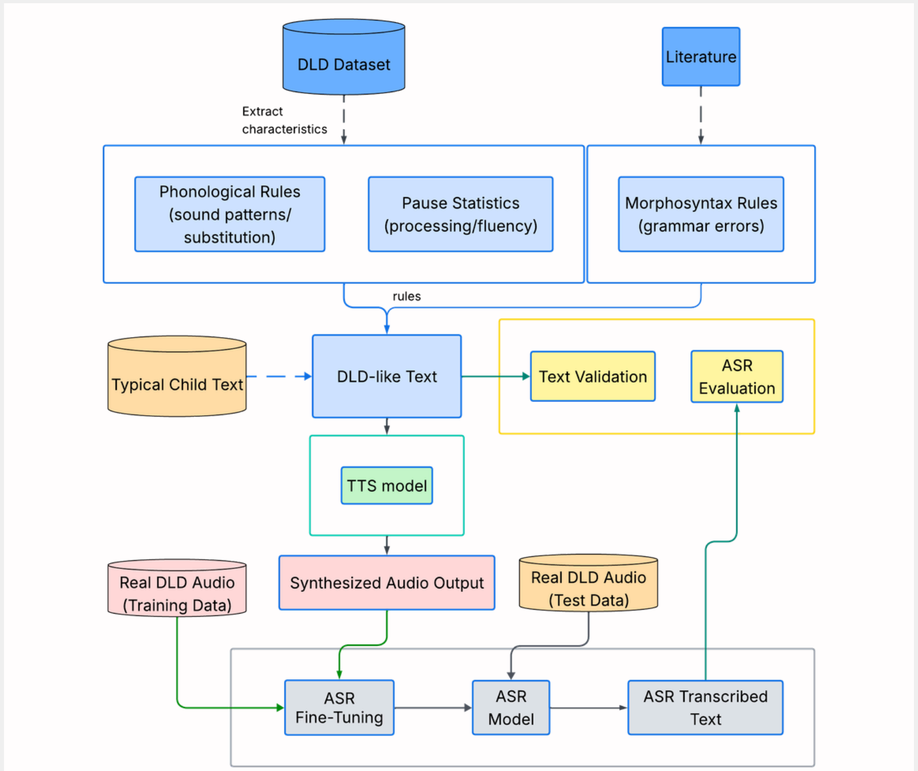
DATA SETS & MODELS

- **CHILDES Zwitserlood** — real Dutch DLD child speech. Train / val / test split (70/10/20%) stratified by speaker, gender, and age.
- **JASMIN corpus** — 15.6 h of typical child read speech, used as source text for the DLD transformation pipeline.
- **XTTS v2 with child reference audio** — 3-second clips from CHILDES used to condition voice cloning. All data handled under strict privacy controls.
- **OpenAI's Whisper large-v3** - ASR model, used for finetuning and evaluation

2. METHODOLOGY

1. Text extraction
2. Morphological errors
3. Phonological errors
4. Disfluency modelling
5. TTS synthesis
6. ASR fine-tuning

(see visual pipeline below)



3. TEXT VALIDATION RESULTS

6.5% WER (train)	94.1% BLEU (train)	8.8% WER (test)	91.6% BLEU (test)
91.5% Combined score (train)	87.1% Overall Feature accuracy(train)	88.5% Combined score (test)	82.6% Overall Feature accuracy(test)

- Pipeline achieves strong lexical and structural alignment with real DLD transcripts
- Phonological generalization on unseen test set (WPhER 31.9%) remains the primary modelling challenge. (Training set WPhER 17.4%)

4. ASR EVALUATION - FILTERED WER (%)

Filtered WER excludes pathological insertion loops (local WER > 100%). Lower = better.

Condition	Score
Zero-shot baseline	36.49%
Fine-tuned on real DLD	27.91%
1 speaker — no DLD	43.05%
1 speaker — DLD	37.32%
1 speaker — DLD severe	41.67%
5 speakers — DLD	38.28%
10 speakers — DLD	37.49%

Legend
- Best real-data result: Fine-tuned on real DLD (27.91%)
- Worst synthetic result: 1 speaker — no DLD (43.05%)

- **Takeaway**:

Synthetic DLD data (best: 37.32%) fails to outperform the zero-shot baseline (36.49%) or real-data fine-tuning (27.91%).

5. FINDINGS

- **Linguistic simulation works.**

DLD-inspired text reduced filtered WER by 5% over error-free synthetic speech ($p < 0.0001$) — confirming linguistic perturbation as a valid design lever.

- **Severity threshold exists**.

Moderate error rates outperform severe perturbation: over-perturbing text disrupts XTTS grapheme-to-phoneme alignment.

- **Acoustic bottleneck persists.**

No synthetic condition outperformed the zero-shot baseline or real-data model — XTTS v2 cannot reliably generate non-standard DLD acoustics.

- **Speaker diversity backfires.**

Scaling to 5 and 10 speakers introduced compounding artifacts and insertion loops, degrading ASR stability.

6. CONCLUSION

- Synthetic data augmentation did not improve ASR performance over the zero-shot baseline.
- The linguistic component of the pipeline is validated.
- The acoustic domain gap is the limiting factor.

7. FUTURE WORK

- Fine-tune TTS on real atypical child speech to close the acoustic gap
- Use mixed synthetic + real training data to prevent catastrophic forgetting
- Add automated quality filters (eMOS thresholds) before training
- Extend pipeline to semantic and pragmatic DLD features
- Generate longer continuous utterances to reduce duration mismatch

wendyhu@tudelft.nl