



Improving Automatic Speech Recognition for Dutch Children with Developmental Language Disorder

Traditional Digital Signal Processing Techniques for Data
Augmentation

Stephanie Gilbert¹

Supervisor(s): Odette Scharenborg¹, YuanYuan Zhang¹

¹EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfillment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Stephanie Gilbert

Final project course: CSE3000 Research Project

Thesis committee: Odette Scharenborg¹, YuanYuan Zhang¹, Jan van Gemert

An electronic version of this thesis is available at <http://repository.tudelft.nl/>

Abstract

This thesis examines whether Dutch adult dysarthric speech can be transformed into speech that is sufficiently child-like to improve Automatic Speech Recognition (ASR) for Dutch children with Developmental Language Disorder (DLD). Due to the limited availability of annotated speech data from children with DLD, existing ASR systems often struggle to accurately recognize this population. To address this data scarcity issue, this work explores the use of traditional Digital Signal Processing (DSP) techniques to augment recordings of Dutch adult speakers with dysarthria to sound more child-like, while preserving overlapping characteristics with DLD.

A series of transformation methods targeting pitch, vocal tract characteristics, spectral properties, and prosodic patterns are implemented and combined into multiple transformation pipelines. The resulting synthetic speech is evaluated through ASR fine-tuning experiments using Whisper-large-v3 and Word Error Rate (WER), as well as an objective speech-quality assessment using Automatic Mean Opinion Score (AMOS) prediction.

The results show that the synthetic child-like speech can indeed improve ASR performance when used as augmentation data. The best-performing configuration combines a targeted subset of mild and moderate hypokinetic and ataxic dysarthric speakers with a prosody-enhanced transformation pipeline, and achieved a WER of 29.00%, outperforming both the zero-shot Whisper-large-v3 baseline (46.48%) and a CHILDES-only fine-tuning baseline (31.57%).

The findings further indicate that source-speaker similarity is more important than dataset size and that conservative prosodic modifications contribute meaningfully to the usefulness of synthetic speech data for ASR fine-tuning. Overall, the results demonstrate that carefully selected dysarthric speech combined with interpretable DSP transformations can provide valuable augmentation data for Dutch children with DLD.

1 Introduction

Automatic Speech Recognition (ASR) systems have improved significantly over the last decade and are now widely used in a variety of applications such as virtual assistants, automatic captioning, accessibility tools, and healthcare technologies [1]. Recent advances in large-scale deep learning models, including transformer-based architectures such as Open AI’s Whisper model [2], have enabled ASR systems to achieve strong performance on several adult speech benchmarks [3]. Despite these developments, one of the main remaining constraints on ASR performance is the limited availability of annotated training data, particularly for underrepresented speaker groups [4]. This problem is especially prevalent for children’s speech. Collecting and annotating children’s speech is difficult and time-consuming because of privacy, ethical, and consent requirements [5]. As a result, publicly available datasets are scarce, and ASR models often receive limited exposure to child speech during training, which contributes to reduced performance on this population [6, 7]. Furthermore, children’s speech is difficult to model because it differs acoustically and linguistically from adult speech. Compared with adults, children tend to have higher fundamental frequency, shorter vocal tracts, less stable articulation, and more variation in pronunciation and speaking rate [8, 7]. Even in clean recordings, these differences make children’s speech difficult to recognize reliably [7].

To address this scarcity, previous research has explored several data augmentation techniques for generating synthetic child-like speech, including Text-to-Speech (TTS), Voice Conversion (VC), and traditional acoustic transformations. Kadyan et al. [9] improved low-resource Punjabi children’s ASR using a two-stage augmentation approach that combined

prosody modification with TTS-based speech synthesis. Similarly, Zhang et al. [10] investigated monolingual and cross-lingual child-to-child voice conversion for Dutch children’s ASR and reported improved recognition performance when the generated speech was used for model fine-tuning. While TTS and VC approaches can generate more natural-sounding synthetic child speech, they often modify several speaker characteristics simultaneously, making it difficult to determine which acoustic properties are responsible for ASR improvements. In contrast, traditional Digital Signal Processing (DSP) transformations are less flexible but more interpretable, therefore better suited for isolating the effects of pitch, spectral structure, and prosody separately. Augmentation techniques such as Vocal Tract Length Perturbation (VTLP) [11] and other acoustic transformations [12] have likewise shown that increasing variability in the training data can improve ASR robustness.

Even so, these efforts have mostly concentrated on typical child speech. For atypical child speech, such as speech produced by children with Developmental Language Disorder (DLD), the data scarcity problem is more severe, and the need for dedicated augmentation strategies remains largely unaddressed. DLD is a neurodevelopmental condition that affects language acquisition and production in children, often leading to atypical pronunciation, inconsistent speech patterns, and greater linguistic variability [13]. Children with DLD may also exhibit reduced intelligibility, atypical prosody, inconsistent articulation, and irregular speech timing, which further complicate ASR recognition [14, 15]. To the best of the author’s knowledge, the CHILDES corpus [16, 17, 18] is currently the only publicly available dataset with Dutch speech recordings from children with DLD that is suitable for ASR training. This apparent scarcity motivates the search for alternative speech sources and data augmentation techniques that can increase the available training data while reducing reliance on large manually collected datasets.

One potential alternative source for generating synthetic training data for Dutch children with DLD is Dutch dysarthric speech. Dysarthria is a motor speech disorder caused by neurological impairment that affects the planning, control, or execution of speech, often resulting in reduced intelligibility, impaired articulation, altered prosody, abnormal speech rate, and respiratory or phonatory instability [19]. Although DLD and dysarthria originate from different underlying disorders, the speech produced in both conditions may share several perceptual and acoustic characteristics relevant to ASR. In particular, both may involve reduced intelligibility, atypical prosody, inconsistent articulation, and irregular timing patterns [19, 15]. The Dutch Corpus of Pathological and Normal Speech (COPAS) contains seven types of dysarthria, ranging in severity from mild to severe [20]. This work hypothesizes that speech from speakers with mild or moderate ataxic or hypokinetic dysarthria may provide the most useful source material for augmentation, as these subtypes are commonly associated with disturbances in articulation, speech timing, rhythm, stress, and overall speech clarity [19]. While no comprehensive study has directly examined the relationship between Dutch child DLD speech and Dutch adult dysarthric speech, the overlap in reported speech characteristics suggests that dysarthric speech may provide a useful starting point for producing child-like speech for ASR training.

Therefore, the main research question is:

To what extent does converting Dutch adult dysarthric speech into Dutch child-like speech using traditional acoustic transformations improve ASR performance for Dutch children with DLD?

To answer this question, synthetic child-like speech is generated from adult dysarthric recordings using a series of interpretable speech-transformation techniques. Different speaker

subsets, transformation pipelines, and training conditions are then used for ASR model fine-tuning which will transcribe recordings from Dutch children with DLD. The generated speech is then assessed using both Word Error Rate (WER) and Automatic Mean Opinion Score (AMOS) prediction.

2 Methodology

2.1 Datasets

2.1.1 CHILDES

The CHILDES corpus [17, 18] provides the target-domain data for this research. It contains longitudinal storytelling recordings from 64 Dutch children with DLD aged 6–10 years, with approximately 180 recordings in total. The recordings were collected using the Dutch Language Test TAK, resulting in connected narrative speech rather than isolated words or read speech. This provides a realistic evaluation setting for ASR systems intended to recognize natural child speech.

2.1.2 COPAS

The augmented datasets were generated from Flemish-Dutch adult dysarthric recordings in the COPAS corpus. COPAS contains recordings from both healthy speakers and speakers with several speech disorders. For this study, only the 87 dysarthric speakers were used.

The corpus contains several recording tasks, including isolated words, sentences, reading passages, and spontaneous speech. For this study, only recordings from the Text (T), *Text Marloes* (TM), sentence repetition (S1 and S2), and spontaneous and semi-spontaneous speech (SS and SSS respectively) tasks were used. These tasks were selected because they are more similar to the CHILDES storytelling recordings and therefore better match the target domain. Isolated-word recordings were excluded because they are more commonly used for dysarthria diagnosis and were not expected to provide substantial benefit for ASR fine-tuning in this setting. Additionally, the SS and SSS recordings do not have manual transcriptions, so an untuned Whisper-large-v3 model was used to generate approximate transcripts. These transcripts may contain recognition errors and may therefore introduce noise into the fine-tuning data.

Although COPAS and CHILDES represent closely related varieties of Dutch, regional pronunciation differences and recording conditions introduce some remaining domain mismatch.

2.1.3 Speaker subset selection

The COPAS corpus contains speakers with a wide range of dysarthria types and severity levels. Three speaker subsets were constructed to compare a narrowly targeted selection with progressively broader and more diverse pools. The first subset (S1, Selected) contains only speakers with mild or moderate hypokinetic or ataxic dysarthria and was expected to provide the closest acoustic match to the target domain. The second subset (S2, Mild/Moderate) contains all labeled types of mild or moderate dysarthria. The third subset (S3, All) includes all dysarthric speakers in the corpus.

The resulting speaker subsets contain 12, 61, and 87 speakers, respectively. Their clinical characteristics are summarized in Table 1. A demographic breakdown of the COPAS dataset can be found in Table 4, as well as the duration of each dataset in Table 5.

Table 1: Clinical characteristics of the COPAS speaker subsets used in this study. Missing type or severity fields are counted as unknown/unlabeled.

Clinical variable	S1 (Selected) (n=12)	S2 (Mild/Moderate) (n=61)	S3 (All Dysarthric) (n=87)
<i>Severity</i>			
Mild	8 (66.7%)	36 (59.0%)	36 (41.4%)
Moderate	4 (33.3%)	25 (41.0%)	25 (28.7%)
Severe	–	–	7 (8.0%)
Unknown/Unlabelled	–	–	19 (21.8%)
<i>Dysarthria Type</i>			
Hypokinetic	9 (75.0%)	9 (14.8%)	16 (18.4%)
Ataxic	3 (25.0%)	3 (4.9%)	5 (5.7%)
Flaccid	–	14 (23.0%)	21 (24.1%)
Spastic	–	6 (9.8%)	8 (9.2%)
Mixed	–	2 (3.3%)	5 (5.7%)
Hyperkinetic	–	3 (4.9%)	3 (3.4%)
UUMND	–	1 (1.6%)	2 (2.3%)
Bulbar	–	–	1 (1.1%)
Hypokinetic?/Mixed?	–	1 (1.6%)	1 (1.1%)
Unknown/Unlabelled	–	22 (36.1%)	25 (28.7%)

2.2 Transformation Methods

The synthetic child-like speech was generated using traditional signal processing techniques designed to increase perceived child-likeness while preserving dysarthric characteristics relevant for augmentation. Each transformation method is implemented independently and targeted a different acoustic difference between adult and child speech, and the transformations were combined into the experimental pipelines described in Section 2.3.

2.2.1 Harmonic Enhancement

Harmonic enhancement was used as a light signal-conditioning step to slightly reinforce the periodic component of the speech signal. It was chosen to reduce breathiness or turbulence without over-smoothing the speech or removing clinically relevant atypical characteristics. The method was implemented with harmonic-percussive source separation (HPSS) via `librosa.effects.harmonic` [21], which extracts a harmonic estimate by separating stable spectral components from transient or noise-like ones [22, 23].

For each recording, the input waveform x was converted to single-precision floating point format and the harmonic component h was extracted with a margin value of 4.0. The enhanced signal was then computed as

$$y = (1 - \lambda)x + \lambda h, \quad (1)$$

with $\lambda = 0.10$. This conservative setting preserved most of the original signal while slightly emphasizing periodic structure.

2.2.2 Adaptive Pitch Modification

Fundamental frequency (F_0) differs strongly between adult and child speech [24, 25], but a fixed pitch shift would ignore speaker-specific variation. To address this, an adaptive pitch modification procedure was implemented.

For each recording, the waveform was converted to 32-bit floating point format and F_0 was estimated using YIN via `librosa.yin` [26, 21]. Invalid values and estimates outside 75–420 Hz were removed, and recordings with fewer than 20 valid pitch frames were left unchanged. The median pitch of the remaining voiced frames was used as the speaker-level estimate:

$$F_{0,\text{med}} = \text{median}(F_0). \quad (2)$$

A conservative child-like target pitch was then selected:

$$F_{0,\text{target}} = \begin{cases} 205 \text{ Hz}, & F_{0,\text{med}} < 150 \text{ Hz}, \\ 220 \text{ Hz}, & 150 \leq F_{0,\text{med}} < 200 \text{ Hz}, \\ 240 \text{ Hz}, & F_{0,\text{med}} \geq 200 \text{ Hz}. \end{cases} \quad (3)$$

These targets were chosen below typical pitch values reported for 6–10-year-old children [24], because full shifting toward those values produced audible artifacts in preliminary listening. The required shift was computed as

$$s = 12 \log_2 \left(\frac{F_{0,\text{target}}}{F_{0,\text{med}}} \right), \quad (4)$$

where s is the pitch shift in semitones and is clipped to $[-0.5, s_{\text{max}}]$, where s_{max} was 1.5 semitones for noisier recordings and 3.0 semitones otherwise, based on spectral flatness. Shifts below 0.6 semitones were not applied. The final transformation used `librosa.effects.pitch_shift` [21], preserving duration while adjusting pitch.

2.2.3 Spectral Envelope Warping

Child speech typically has higher vocal-tract resonance frequencies than adult speech [25]. To approximate this difference without altering the full spectrum, spectral envelope warping was applied to modify the broad spectral shape while preserving harmonic detail.

The signal was transformed using the short-time Fourier transform with an FFT size of 2048 and a hop length of 512 samples [21]. The logarithmic magnitude spectrum was smoothed along the frequency axis with a one-dimensional Gaussian filter ($\sigma = 8$) [27] to estimate the spectral envelope. This envelope was then warped with a formant scaling ratio r_f . Let $E(f)$ denote the smoothed spectral envelope and r_f the formant-scaling ratio. The warped envelope $E'(f)$ was computed as:

$$E'(f) = E\left(\frac{f}{r_f}\right). \quad (5)$$

A value of $r_f = 1.18$ was used, corresponding to an approximate 18% upward shift of the envelope. Rather than replacing the original spectrum, the warped envelope was converted

into a frequency-dependent gain function and applied to the original magnitude spectrum, with the gain limited to ± 8 dB. The transformed signal was finally mixed with the original recording as

$$y = (1 - \lambda)x + \lambda x', \quad (6)$$

with $\lambda = 0.60$, retaining more of the warped signal while keeping some of the original speech characteristics.

2.2.4 Spectral Tilt Modification

Child speech is often described as having a brighter spectral profile than adult speech, with relatively more high-frequency energy. To approximate this, spectral tilt was modified using a first-order pre-emphasis filter [28].

The filter was implemented as

$$y[n] = x[n] - \alpha x[n - 1], \quad (7)$$

with $\alpha = 0.90$. Because applying the filter directly produced speech that sounded unnaturally sharp, the filtered signal was blended with the original recording:

$$z[n] = (1 - \lambda)x[n] + \lambda y[n], \quad (8)$$

with $\lambda = 0.18$. This produced a modest increase in high-frequency emphasis while avoiding excessive brightness.

2.2.5 Phrase-Level Prosody Modification

Child speech often differs from adult speech in pitch variability, stress, and intonation [29]. To approximate these suprasegmental patterns, a phrase-level prosody modification was applied to selected regions within each utterance.

Candidate regions were detected using Librosa onset detection. An onset strength envelope was computed, prominent onsets were ranked, and up to five regions were selected with at least 550 ms spacing. Each region was approximately 650 ms long. Within each region, pitch was increased by 0.3–1.2 semitones using `librosa.effects.pitch_shift`, while local speaking rate was reduced by up to 10% using `librosa.effects.time_stretch`. Amplitude stress was modeled using a smooth sinusoidal weighting function,

$$w(t) = \sin(\pi t)^{0.7}, \quad (9)$$

with an overall gain of 1.0–4.5 dB. Modified regions were blended with the original signal using 40 ms crossfades.

2.2.6 Praat-Based Childification

As an alternative to the custom transformations, childification was also implemented with Praat through Parselmouth [30, 31]. Unlike the other methods, this approach uses Praat’s **Change Gender** function to jointly modify fundamental frequency and formant structure in a single source-filter transformation.

For each recording, the median pitch was estimated with a search range of 75–600 Hz and scaled by $r_p = 1.20$:

$$F_{0,\text{target}} = r_p \cdot F_{0,\text{med}}. \quad (10)$$

The **Change Gender** parameters were set to a formant-shift ratio of 1.20, a pitch-shift ratio of 1.20, and a duration factor of 1.0. This configuration was chosen to approximate the shorter vocal tract and higher pitch of school-aged children while preserving temporal structure. The method was included to test whether a more integrated, physiologically motivated transformation could produce more natural child-like speech than the modular DSP-based pipeline.

2.3 Transformation Pipelines

The individual transformations described in the previous section were combined into three transformation pipelines. Rather than evaluating every transformation independently, the pipelines were designed to compare three distinct childification strategies for generating synthetic training data: a modular transformation approach, a prosody-enhanced variant, and an integrated speech-transformation method. The resulting comparison aims to determine which strategy produces the most useful speech augmentation techniques for improving ASR performance on Dutch children with DLD.

2.3.1 P1: Modular Childification Pipeline

The first pipeline consists of harmonic enhancement, adaptive pitch modification, spectral envelope warping, and spectral tilt modification, from Section 2.2. This pipeline was designed to evaluate whether a combination of targeted acoustic modifications is sufficient to reduce the mismatch between adult dysarthric speech and child speech. By focusing primarily on vocal source and vocal-tract characteristics while avoiding explicit prosodic manipulation, P1 serves as a controlled baseline for assessing the contribution of more complex transformations.

2.3.2 P2: Prosody-Enhanced Childification Pipeline

The second pipeline extends P1 with phrase-level prosody modification targeting stress, rhythm, speaking rate, and intonation, as described in Section 2.2.5. P2 therefore evaluates whether incorporating these suprasegmental characteristics provides additional benefits for synthetic speech augmentation beyond the segmental modifications applied in P1. Because atypical prosody may also contain characteristics relevant to DLD, the modifications were intentionally conservative to avoid over-processing the speech signal.

2.3.3 P3: Praat-Based Childification Pipeline

The third pipeline consists solely of the Praat-based childification method described in Section 2.2.6. Unlike P1 and P2, which combine multiple independent transformations, this approach jointly modifies pitch and formant structure within a single source-filter framework.

P3 was included as an alternative childification strategy that relies on a widely used speech-processing method rather than a collection of custom transformations. This enables a comparison between a highly controllable modular approach (P1 and P2) and a more integrated transformation method, helping determine whether increased control over individual speech characteristics translates into improved ASR performance.

2.4 Experimental Design

The experiments are designed to investigate how speaker selection, transformation complexity and data quantity influence the usefulness of the generated data when used for ASR fine-tuning. Each system is evaluated against two baselines: the off-the-shelf Whisper-large-v3 model and a Whisper-large-v3 model fine-tuned only on the CHILDES dataset.

The experiments evaluate three experimental factors. The first factor consists of the three COPAS speaker subsets (S1–S3) described in Section 2.1.3. The second factor consists of the three transformation pipelines (P1–P3) described in Section 2.3. The third experimental factor consists of two training conditions. The first condition uses only the generated synthetic speech (T1). This setting investigates whether the generated recordings could provide a viable alternative training source in scenarios where little or no speech from low-resource domains such as children with DLD is available. The second condition combines the generated synthetic speech with the CHILDES training set (T2). This setting investigates whether the synthetic data can improve ASR performance when used to augment an existing DLD dataset. Together, T1 and T2 evaluate both the replacement and augmentation potential of the generated speech, resulting in eighteen total experimental conditions.

2.4.1 ASR Fine-tuning

The CHILDES DLD dataset is divided into training, validation, and test sets using a speaker-independent split. Approximately 70% of the recordings are used for training, 10% for validation, and 20% for testing. The validation and test sets remain unchanged throughout all experiments to ensure a fair comparison between training conditions.

All ASR experiments are conducted using the SpeechBrain toolkit [32]. The Whisper-large-v3 model [2] is fine-tuned using Low-Rank Adaptation (LoRA) [33], a parameter-efficient fine-tuning method that updates only a small subset of model parameters. Models are trained for up to 50 epochs on an NVIDIA A40 GPU. Following the standard SpeechBrain fine-tuning procedure, model checkpoints are evaluated on the validation set after each epoch, and the checkpoint achieving the lowest validation-set Word Error Rate (WER) is selected as the final model for testing.

ASR performance from each experiment is evaluated using Word Error Rate (WER), a standard metric used to evaluate the quality of transcriptions produced by ASR systems [28]. WER measures the number of word-level errors produced by an ASR system relative to a reference transcription and therefore provides a direct measure of transcription accuracy. Specifically, WER is based on the minimum number of substitutions, deletions, and insertions required to transform the recognized transcription into the reference transcription. Lower WER values indicate better recognition performance.

WER is calculated as:

$$\text{WER} = \frac{S + D + I}{N} \times 100 \quad (11)$$

where (S) denotes the number of substitutions, (D) the number of deletions, (I) the number of insertions, and (N) the total number of words in the reference transcription.

2.4.2 Speech Quality Evaluation

In addition to ASR performance, the generated speech is evaluated using Automatic Mean Opinion Score (AMOS) prediction. Speech quality is evaluated for each experiment and

both baselines using MBNet [34], a neural network trained to predict Mean Opinion Score (MOS) values directly from speech recordings. The AMOS score reported for a dataset is calculated as the average predicted score across all recordings:

$$AMOS = \frac{1}{N} \sum_{i=1}^N m_i \quad (12)$$

where m_i denotes the predicted score for recording i , and N is the total number of recordings in the dataset. Higher AMOS values indicate higher predicted speech quality.

AMOS is used as a secondary evaluation metric alongside WER to investigate whether transformations that improve the perceptual quality of the generated speech also improve ASR performance. Since MBNet was trained on typical speech and was not specifically developed for dysarthric speech or speech produced by children with DLD, the resulting scores should be interpreted primarily as relative quality indicators for comparing experimental conditions rather than as absolute measures of perceptual quality.

3 Results

3.1 ASR Performance

Table 2 presents the WER for each speaker subset, transformation pipeline, and training condition. The baseline systems achieved WERs of 46.48% for the zero-shot Whisper-large-v3 model and 31.57% for the CHILDES-only fine-tuned model. The best-performing configuration was S1-P2-T2, with a WER of 29.00%, improving on both baselines. In total, 17 of the 18 evaluated configurations outperformed the zero-shot baseline.

3.1.1 Effect of Speaker Subset Selection

S1 consistently yielded the best results across pipelines and training conditions. Its best WER was 29.00%, compared to 34.62% for S2 and 33.96% for S3. This pattern was especially clear under T2, where S1 outperformed S2 and S3 for both P1 and P2. Although S2 and S3 included more speakers, neither matched the best S1 result.

3.1.2 Effect of Transformation Pipeline

Table 6 summarizes the average WER for each pipeline. Under T1, P3 performed best on average with 39.29%, closely followed by P2 with 39.86%, while P1 performed worst with 41.34%. Under T2, P2 performed best with 33.99%, ahead of P3 at 35.93% and P1 at 42.56%. Adding the prosody stage in P2 improved performance across all three speaker subsets under T2. P3 was competitive, especially for S3, but did not surpass the best P2 result.

3.1.3 Effect of Training Condition

Adding CHILDES training data improved performance for P2 and P3. For P2, the average WER decreased from 39.86% under T1 to 33.99% under T2. For P3, it decreased from 39.29% to 35.93%. At the individual level, T2 was better than T1 in eight of the nine configurations. The largest gain was for S1-P2, where WER dropped from 39.51% to 29.00%. The only degradation occurred for S2-P1, where WER increased from 42.40% to 51.16%.

Overall, combining synthetic speech with CHILDES data generally improved ASR performance, with the strongest gains for the P2 pipeline.

3.1.4 Relationship Between AMOS and WER

To assess whether predicted speech quality is related to ASR performance, Pearson and Spearman correlations were computed between AMOS and the corresponding T1 WER values. A moderate negative correlation was observed (Pearson $r = -0.56$, Spearman $\rho = -0.58$), suggesting that higher AMOS scores were generally associated with lower WER. However, neither correlation was statistically significant ($p > 0.05$), likely due to the limited number of evaluated configurations.

AMOS should therefore be interpreted cautiously in this setting. As shown in Table 3, the P2 configurations consistently produced AMOS scores closer to the CHILDES baseline than the P1 configurations, while the highest-scoring P3 configurations were substantially closer to the COPAS baseline. This indicates that maximizing predicted speech quality alone is not sufficient to produce the most useful synthetic training data.

Overall, the results suggest that closeness to the CHILDES baseline may be more informative than absolute AMOS when evaluating childification strategies for this task.

Table 2: Test-set WER (%) for each speaker subset and transformation pipeline. Lower values indicate better ASR performance. T1 uses only synthetic speech for fine-tuning, while T2 combines synthetic speech with the CHILDES training set.

Subset	P1		P2		P3	
	T1	T2	T1	T2	T1	T2
S1	39.71	34.82	39.51	29.00	42.67	34.69
S2	42.40	51.16	38.01	34.62	37.73	39.15
S3	41.91	41.70	42.07	38.34	37.46	33.96

4 Discussion

4.1 Does Synthetic Childification Improve ASR Performance?

The results indicate that transforming adult dysarthric speech into child-like speech can improve ASR performance for Dutch children with DLD, particularly when used as augmentation data. Although most configurations improved upon the zero-shot baseline, only S1-P2-T2 outperformed the CHILDES-only baseline, suggesting that the generated speech captures useful characteristics without fully replicating the variability of real child DLD speech.

These findings are consistent with prior work showing that synthetic child speech can support ASR improvement. Kadyan et al. [9] reported gains using TTS-based augmentation, Zhang et al. [10] found voice conversion effective for child-like training data, and Kathania et al. [35] showed that spectral warping can also improve recognition. The present study leads to a similar conclusion: synthetic child-like speech can be a useful supplement for low-resource child ASR. However, unlike the TTS-based approach investigated by Kadyan et al. [9] and the voice-conversion approaches explored by Zhang et al. [10], the present study relies exclusively on interpretable DSP transformations. While TTS and VC methods

Table 3: AMOS results for the baseline datasets and generated speech across speaker subsets and transformation pipelines. Δ_{CHILDES} and Δ_{COPAS} indicate the absolute difference from the corresponding baseline mean AMOS score.

Dataset	Mean AMOS	Std. Dev.	Δ_{CHILDES}	Δ_{COPAS}
CHILDES (all)	1.590	0.355	0.000	0.232
COPAS (all)	1.822	0.523	0.232	0.000
S1-P1	1.508	0.241	0.082	0.314
S1-P2	1.526	0.292	0.064	0.296
S1-P3	1.664	0.244	0.074	0.158
S2-P1	1.488	0.235	0.102	0.334
S2-P2	1.511	0.281	0.079	0.311
S2-P3	1.805	0.585	0.215	0.017
S3-P1	1.473	0.188	0.117	0.349
S3-P2	1.497	0.248	0.093	0.325
S3-P3	1.847	0.589	0.257	0.025

can generate highly realistic child-like speech, they simultaneously modify multiple acoustic characteristics, making it difficult to isolate the contribution of individual factors. In contrast, the modular design of the proposed pipeline allows pitch, spectral, and prosodic modifications to be evaluated separately. The observed improvements therefore provide evidence that relatively simple and explainable acoustic transformations can already produce useful augmentation data for DLD ASR.

Overall, the findings suggest that synthetic dysarthric speech is a promising augmentation source for low-resource DLD ASR, but not yet a replacement for real child recordings. The strongest performance was achieved when synthetic and real child speech were combined.

4.2 Importance of Domain Similarity

A notable finding of this study is that the strongest ASR performance was obtained using the most targeted speaker subset rather than the largest speaker subset. The S1 subset was specifically designed to contain speakers whose dysarthric characteristics were expected to be most similar to those observed in children with DLD, while S2 and S3 progressively increased speaker diversity and dataset size (Table 1).

These results suggest that the usefulness of synthetic training data depends not only on the amount of available data, but also on the similarity between the source and target domains. Although larger datasets are often associated with improved machine learning performance, the results of this study indicate that increasing speaker diversity alone does not necessarily improve ASR performance for atypical speech recognition tasks.

These results indicate that within the tested COPAS subsets, domain similarity appeared to be more important than dataset size when selecting source speakers for synthetic data generation. This underlines the importance of careful speaker selection when using pathological speech as a basis for child speech augmentation.

4.3 Importance of Prosodic Modification

A second notable finding is the strong performance of the P2 transformation pipeline. Compared to P1, P2 introduced additional phrase-level prosodic modifications while retaining the same pitch and spectral transformations. The consistent improvements observed for P2, as shown in Table 2 suggest that prosodic characteristics play an important role when generating synthetic child-like speech for ASR training.

The prosodic modifications used in this study were intentionally conservative in order to preserve potentially useful dysarthric characteristics while introducing child-like speech patterns. The results suggest that this balance may be important when designing transformation pipelines for atypical speech augmentation.

4.4 Speech Quality versus ASR Utility

AMOS and ASR utility appear related, but they do not have the same objective. Although a moderate negative correlation was observed between AMOS and WER, the highest AMOS scores did not coincide with the best ASR performance: P2 achieved the lowest WER, while P3 achieved the highest AMOS scores.

One explanation is that MBNet was trained to estimate perceptual quality for typical speech [34], whereas this study focuses on atypical speech from children. As a result, improvements in predicted naturalness do not necessarily imply greater similarity to the target domain.

This is also reflected in the baseline comparisons. P3 produced AMOS scores closer to the COPAS recordings, while P2 was closer to the CHILDES baseline and yielded better ASR performance. These results suggest that target-domain similarity may be more informative than maximizing AMOS alone.

Overall, AMOS is useful as a relative quality indicator, but it does not directly capture whether synthetic speech is suitable for atypical ASR training.

5 Responsible Research

5.1 Inclusivity

ASR systems typically perform worse on children’s speech and other atypical speech than on typical adult speech, which can limit access to speech-based technologies for underrepresented groups. This research aims to support more inclusive ASR by improving recognition for Dutch children with DLD, a population for which limited training data is available.

At the same time, the generated speech should not be treated as a replacement for recordings from the target population. Further validation on larger and more diverse datasets is needed before such systems could be used in practice.

5.2 Ethics and Privacy

This study uses the publicly available CHILDES and COPAS datasets and does not involve new data collection. Synthetic speech should be viewed as a complement to, rather than a replacement for, representative recordings from the target population.

5.3 Reproducibility

Reproducibility was considered throughout the project. Transformation parameters, dataset selection procedures, and training configurations are documented in this thesis, and the implementation relies on widely used open-source tools including Whisper, SpeechBrain, Praat, Parselmouth, and librosa.

5.4 Use of Generative AI

In accordance with the TU Delft GenAI guidelines for research papers, this section discloses the use of generative AI tools, primarily ChatGPT by OpenAI, during this work. These tools were used in various bounded ways throughout the project.

First, AI was used as programming assistance during development of the speech transformation pipeline. This included discussing initial digital signal processing approaches and debugging Python code.

Second, AI was used as a research and writing aid. This included discussing the structure of the thesis, improving clarity and readability, formatting results tables, and reference formatting.

No AI-generated text was included without review and modification. All factual claims, implementation details, experimental settings, and dataset descriptions were verified against primary sources, published literature, project code, or the original datasets. No sensitive, personal, or confidential data was entered into any AI system, and full responsibility for the research and final thesis remains with the author.

6 Limitations and Future Work

Several limitations should be considered when interpreting these results. First, the experiments were conducted on a single Dutch DLD corpus (CHILDES) using a single train/validation/test split. Consequently, it remains unclear how well the findings generalize to other datasets, age groups, languages, or speech disorders, and the reported WER values may be influenced by the specific speaker distribution in the selected split. Future work should therefore evaluate the approach on larger datasets and multiple data splits.

A second limitation concerns the COPAS corpus. The SS and SSS recordings do not contain manual transcriptions, requiring the use of Whisper-generated transcripts that may contain recognition errors. In addition, regional and recording-condition differences between the Flemish Dutch COPAS recordings and the Dutch CHILDES recordings introduce domain mismatch. Future work should investigate manually transcribed corpora and domain-adaptation techniques to reduce these effects.

Finally, this study relied exclusively on objective AMOS predictions and interpretable DSP-based transformations. Human evaluation would provide a more direct assessment of child-likeness and naturalness, while comparisons with voice-conversion and text-to-speech approaches would help determine how competitive DSP-based augmentation remains relative to more advanced speech-generation methods.

7 Conclusion

This thesis investigated whether adult dysarthric speech can be transformed into child-like speech to improve ASR for Dutch children with DLD. To address this question, a set of

interpretable digital signal processing transformations was applied to recordings from the COPAS dysarthric speech corpus. The resulting synthetic speech was used to fine-tune Whisper-large-v3 under different speaker selection, transformation pipeline, and training conditions.

The results show that synthetic child-like speech can improve ASR performance when used as augmentation data. The best-performing configuration, S1-P2-T2, achieved the lowest WER and was the only configuration to outperform both the zero-shot and CHILDES-only baselines. The experiments further demonstrated that carefully selected source speakers produced better results than larger and more diverse speaker sets, highlighting the importance of domain similarity. In addition, the strong performance of the prosody-enhanced P2 pipeline suggests that prosodic characteristics contribute meaningfully to the generation of useful synthetic training data for children.

Overall, the findings demonstrate that carefully selected dysarthric speech combined with interpretable transformations can provide valuable augmentation data for low-resource DLD ASR. The results further suggest that similarity to the target population is more important than maximizing either dataset size or predicted speech quality. While additional work is required to validate the approach on larger datasets and through human evaluation, the proposed methodology represents a promising step towards improving ASR accessibility for children with DLD.

References

- [1] H. Ahlawat, N. Aggarwal, and D. Gupta, “Automatic speech recognition: A survey of deep learning techniques and approaches,” *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 201–237, 2025.
- [2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022, openAI technical report. [Online]. Available: <https://cdn.openai.com/papers/whisper.pdf>
- [3] J. Li, “Recent advances in end-to-end automatic speech recognition,” *APSIPA Transactions on Signal and Information Processing*, vol. 11, p. e8, 2022.
- [4] A. Xiao, W. Zheng, G. Keren, D. Le, F. Zhang, C. Fuegen, O. Kalinli, Y. Saraf, and A. Mohamed, “Scaling asr improves zero and few shot learning,” in *Proceedings of Interspeech 2022*. Incheon, Korea: ISCA, Sep. 2022, 18–22 September 2022.
- [5] Z. Yue, M. Barberis, T. Patel, J. Dineley, W. Doedens, L. Stipdonk, Y. Zhang, E. de Witte, E. Loweimi, H. V. hamme, D. Satoer, M. Ruiter, L. M. Velazquez, N. Cummins, and O. Scharenborg, “Challenges and practical guidelines for atypical speech data collection, annotation, usage and sharing: A multi-project perspective,” in *Proceedings of Interspeech 2025*, 2025, pp. 3943–3947.
- [6] D. Liu, J. Yang, I. Albrecht-Buehler, H. Qin, S. Li, Y. Hu, A. Nassereldine, and J. Xiong, “Automatic screening for children with speech disorder using automatic speech recognition: Opportunities and challenges,” in *Proceedings of the AAAI Symposium Series*, vol. 4, no. 1, Nov. 2024, pp. 308–313.
- [7] R. Sobti, K. Guleria, and V. Kadyan, “Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges,” *Multimedia Tools and Applications*, vol. 83, no. 35, pp. 81 933–81 995, 2024.
- [8] V. Bhardwaj, M. D. Troia, G. Stampone, A. Mantri, and A. Kumar, “Automatic speech recognition (asr) systems for children: A systematic literature review,” *Applied Sciences*, vol. 12, no. 9, p. 4419, 2022.
- [9] V. Kadyan, H. Kathania, P. Govil, and M. Kurimo, “Synthesis speech based data augmentation for low resource children asr,” in *Speech and Computer*. Springer, 2021, pp. 317–326.
- [10] Y. Zhang, Z. Yue, T. Patel, and O. Scharenborg, “Improving child speech recognition with augmented child-like speech,” in *Proceedings of Interspeech 2024*, 2024, pp. 5183–5187.
- [11] N. Jaitly and G. Hinton, “Vocal tract length perturbation (vtlp) improves speech recognition,” *Proceedings of ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013.
- [12] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation for speech recognition,” *Proceedings of Interspeech*, 2015.

- [13] D. V. M. Bishop, M. J. Snowling, P. A. Thompson, and T. Greenhalgh, “Phase 2 of catalise: A multinational and multidisciplinary delphi consensus study of problems with language development,” *Journal of Child Psychology and Psychiatry*, vol. 58, no. 10, pp. 1068–1080, 2017.
- [14] L. Pennington and D. V. M. Bishop, “Speech, language, and communication difficulties in children with developmental language disorders,” *Current Paediatrics*, vol. 19, no. 3, pp. 157–163, 2009.
- [15] D. Rosero *et al.*, “Synthetic child speech generation for improving automatic speech recognition of children with speech sound disorders,” in *Proceedings of Interspeech*, 2024.
- [16] B. MacWhinney, “The childes project: Tools for analyzing talk,” 2000.
- [17] R. Zwitterlood, “Language growth in dutch school-age children with specific language impairment,” Ph.D. dissertation, Utrecht University, 2014. [Online]. Available: https://www.lotpublications.nl/Documents/356_fulltext.pdf
- [18] R. Zwitterlood, M. van Weerdenburg, L. Verhoeven, and F. Wijnen, “Development of morphosyntactic accuracy and grammatical complexity in dutch school-age children with specific language impairment,” *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 3, pp. 891–905, 2015.
- [19] J. R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, 4th ed. Elsevier, 2020.
- [20] G. V. Nuffelen, M. D. Bodt, H. V. Hamme, P. Koolagudi, and J. Verhelst, “The dutch corpus of pathological and normal speech (copas),” in *Proceedings of Interspeech*, 2009, pp. 2895–2898.
- [21] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python,” in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–25.
- [22] D. FitzGerald, “Harmonic/percussive separation using median filtering,” in *Proceedings of the 13th International Conference on Digital Audio Effects*, 2010, pp. 246–253.
- [23] J. Driedger, M. Müller, and S. Disch, “Extending harmonic-percussive separation of audio signals,” in *Proceedings of the 15th International Society for Music Information Retrieval Conference*, 2014, pp. 611–616.
- [24] D. N. Sorenson and Y. Horii, “A fundamental frequency investigation of children ages 6 through 10,” *Journal of Speech and Hearing Research*, vol. 32, no. 1, pp. 197–201, 1989.
- [25] S. Bennett, “Vowel formant frequency characteristics of preadolescent males and females,” *Journal of the Acoustical Society of America*, vol. 69, no. 1, pp. 231–238, 1981.
- [26] A. de Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.
- [27] P. Virtanen *et al.*, “Scipy 1.0: Fundamental algorithms for scientific computing in python,” *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.

- [28] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.
- [29] J. E. Kallay, M. A. Redford, and S. R. Speer, “Prosodic development during the early school-age years,” *Journal of Speech, Language, and Hearing Research*, vol. 66, no. 2, pp. 506–522, Feb. 2023.
- [30] P. Boersma and D. Weenink, “Praat: Doing phonetics by computer,” 2025, version 6.4. [Online]. Available: <https://www.praat.org>
- [31] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing parselmouth: A python interface to praat,” *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.
- [32] M. Ravanelli *et al.*, “Speechbrain: A general-purpose speech toolkit,” in *ICASSP*, 2021.
- [33] E. J. Hu *et al.*, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
- [34] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, “Mbnet: Mos prediction for synthesized speech with mean-bias network,” *arXiv preprint arXiv:2103.00110*, 2021.
- [35] H. Kathania, V. Kadyan, S. V. G. Kadiri, and M. Kurimo, “Data augmentation using spectral warping for low resource children asr,” *Journal of Signal Processing Systems*, Nov. 2022.

A COPAS Speaker Demographic Statistics

Table 4: Demographic characteristics of the COPAS speaker subsets used in this study.

Demographic variable	S1 (Selected) (n=12)	S2 (Mild/Moderate) (n=61)	S3 (All Dysarthric) (n=87)
<i>Gender</i>			
Male	7 (58.3%)	36 (59.0%)	52 (59.8%)
Female	4 (33.3%)	24 (39.3%)	34 (39.1%)
Unknown	1 (8.3%)	1 (1.6%)	1 (1.1%)
<i>Age Group</i>			
<18 years	2 (16.7%)	8 (13.1%)	13 (14.9%)
18–60 years	2 (16.7%)	35 (57.4%)	47 (54.0%)
60+ years	7 (58.3%)	17 (27.9%)	25 (28.7%)
Unknown	1 (8.3%)	1 (1.6%)	2 (2.3%)

B Characteristics of Speech Datasets

Table 5: Characteristics of the speech datasets used in the experiments.

Dataset	Files	Minutes	Hours
CHILDES (all)	314	128.82	2.15
COPAS (all)	782	407.47	6.79
S1	62	22.13	0.37
S2	198	55.99	0.93
S3	406	113.68	1.89

C Additional ASR Results

Table 6: Average WER (%) across all speaker subsets for each transformation pipeline.

Pipeline	Mean T1 WER	Mean T2 WER
P1	41.34	42.56
P2	39.86	33.99
P3	39.29	35.93

Table 7: Difference in WER (%) relative to the Whisper-large-v3 zero-shot baseline (46.48%). Negative values indicate an improvement over the baseline, while positive values indicate worse performance.

Subset	P1		P2		P3	
	T1	T2	T1	T2	T1	T2
S1	-6.77	-11.66	-6.97	-17.48	-3.81	-11.79
S2	-4.08	+4.68	-8.47	-11.86	-8.75	-7.33
S3	-4.57	-4.78	-4.41	-8.14	-9.02	-12.52

Table 8: Difference in WER (%) relative to the CHILDES-only fine-tuning baseline (31.57%). Negative values indicate an improvement over the baseline, while positive values indicate worse performance.

Subset	P1		P2		P3	
	T1	T2	T1	T2	T1	T2
S1	+8.14	+3.25	+7.94	-2.57	+11.1	+3.12
S2	+10.83	+19.59	+6.44	+3.05	+6.16	+7.58
S3	+10.34	+10.13	+10.50	+6.77	+5.89	+2.39