



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

Julie van Montfrans¹

Supervisors: Odette Scharenborg¹, Yuanyuan Zhang¹

¹EEMCS, Delft University of Technology, The Netherlands

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

Name of the student: Julie van Montfrans
Final project course: CSE3000 Research Project
Thesis committee: Odette Scharenborg, Jan van Gemert

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

Abstract

Automatic Speech Recognition (ASR) systems perform poorly on speech of children with Developmental Language Disorder (DLD) due to a scarcity of training data and the high variability in speech of these children. This study investigates whether speech created through cross-lingual child-to-child voice conversion (VC) can improve ASR performance for Dutch children with DLD. To this end, multiple data augmentation pipelines were developed using the MeanVC zero-shot VC model, generating synthetic child speech with different source-target speaker pairing strategies (most similar, random, least similar) and varying augmentation sizes. The augmented datasets were used to fine-tune a Whisper-based ASR model.

The results show that fine-tuning the ASR on a combination of synthetic speech and source speech improved ASR performance over an out-of-the-box baseline, but did not lead to better performance than fine-tuning it on real speech of children with DLD alone. Interestingly, speaker-dissimilar VC pairings yielded the best ASR performance. Subjective Mean Opinion Score (MOS) evaluations suggest that synthetic speech was perceived as less natural, less intelligible, and less child-like than real speech and that AMOS predictions only partially align with human judgments and ASR performance. These findings suggest that while cross-lingual voice-converted speech can have value as supplementary training data, improvements in speech quality and voice conversion methods are needed before it can serve as an effective replacement for real speech of a child with DLD in ASR training.

1 Introduction

Automatic speech recognition (ASR) systems are widely used in a variety of applications and have undergone rapid development in recent years. Despite these advances, processing children’s speech remains a challenge [1]. This difficulty is pronounced for children with developmental language disorder (DLD) [2]. DLD is a condition that impacts the language development of children without being caused by hearing impairment or cognitive deficits [3]. The language of children with DLD may be sparse, nonfluent, less intelligible and agrammatic [4]. This makes communicating and learning much more difficult for them. Reliable ASR systems could provide valuable support for these children by improving accessibility in communication and learning environments.

Previous research has identified two main reasons why ASR systems perform significantly worse for children with DLD than on typical adult speech. First, there is a profound mismatch between the data used to train ASR systems, which is mostly typical adult speech, and the speech of children with DLD, which often has different phonetic properties [2]. Training or fine-tuning the ASR systems on more child- and DLD-like speech could improve ASR performance on this kind of speech. However, this solution is hindered in practice by a second problem: there is a severe scarcity of publicly available datasets containing speech from Dutch-speaking children with DLD; to my knowledge, the only publicly available dataset contains fewer than five hours of speech.

To mitigate this issue, different techniques have been explored, one of them being data augmentation. Wan (2025) investigated the effect of using speed perturbation (SP) and vocal tract length perturbation (VTLP) to perform data augmentation and showed that these methods clearly improve the performance of ASR for speech of children with DLD.

Furthermore, voice conversion (VC) has emerged as a powerful tool for generating synthetic speech data. VC is a technique that transforms a source speaker’s voice into that of a target speaker while preserving the linguistic content of the original utterance [5].

Therefore, it can be used to create new speech data. In recent years, VC has seen a surge in development, transitioning from statistical mappings to powerful neural and generative models, under which variational autoencoder-based systems and generative adversarial networks (GANs) like the CycleGAN-VC [6] and StarGAN-VC [7]. These architectures have created the possibility for non-parallel and multi-domain VC. Non-parallel VC refers to methods that learn mappings between source and target speech without requiring parallel data, while multi-domain VC is a technique that learns mappings across multiple target domains [8]. Moreover, the Voice Conversion Challenge 2020 [9] demonstrated that modern deep learning based VC-systems can achieve high speaker similarity between the synthesized speech and the target speech, even in cross-lingual settings. More recently, models like MeanVC [10] have introduced lightweight, streaming zero-shot voice conversion based on diffusion transformer architectures. MeanVC incorporates mean flows, which regress the average velocity field in a flow-matching framework, enabling efficient one-step (1-NFE) generation.

While most existing research employs adult-to-child VC, this approach may be suboptimal due to fundamental acoustic differences between adult and child speech. For instance, child speech typically exhibits a higher fundamental frequency, longer and more variable segmental durations, higher formant frequencies and a larger variability in both temporal and spectral characteristics within a speaker [11]. With child-to-child VC these acoustic traits are kept. Zhang et al. (2024) recently showed that using child-to-child cross-lingual VC significantly improved child ASR performance. The use of cross-lingual data as target speech mitigates the issue of data scarcity, as more child speech data is available in other languages. Moreover, the linguistic deficiencies of the DLD speech are retained, since VC preserves the content of the source speaker. This makes cross-lingual VC a suitable technique for synthesizing DLD- and child-like speech.

This study therefore investigates to what extent cross-lingual child-to-child VC using child speaker information from non-Dutch children’s speech without DLD can improve ASR performance for Dutch children with DLD. This research question is broken down into the following research questions (RQ):

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?

RQ1 is most directly related to the primary research question, as it evaluates the actual effect of the use of the synthesized data on ASR performance for speech produced by Dutch children with DLD. RQ2 and RQ3 evaluate the naturalness and quality of the generated speech using MOS and AMOS. MOS is a subjective evaluation metric in which human participants rate speech quality, whereas AMOS is an automatic prediction model which estimates these MOS ratings. The combination of these evaluation metrics allows us to examine whether these quality assessments align, and whether synthesized speech of higher perceptual quality is associated with lower word error rates (WER). Moreover, prior work suggests that source-target speaker similarity can influence VC performance [5]. Therefore, RQ4 investigates which speaker pairing strategy yields the best quality speech. Together, these analyses provide insight into the conditions under which fine-tuning on synthetic data

created through VC can most effectively improve ASR performance.

2 Methodology

In order to investigate the extent to which synthesized speech data has a positive influence on ASR, we fine-tuned an ASR model (see section 2.2) on this speech to see whether it improves ASR performance. The different experiments performed with the ASR are explained in section 2.3. In order to create the synthesized data, source and target data were used (see section 2.1) in VC (see section 2.5). For the VC, three different strategies based on speaker similarity are used. To calculate this similarity, speaker embeddings were created as described in 2.4.1. How the source-target pairs were created and how we decided which utterance per speaker to use for the conversion are explained in section 2.4.2. and 2.4.3 respectively. Section 2.6 describes the metrics used to evaluate the synthesized speech.

2.1 Datasets

2.1.1 CHILDES Zwitserlood

The CHILDES Zwitserlood corpus contains speech recordings from 60 Dutch monolingual children with DLD (aged 6-10 years) and 90 typically developing Dutch monolingual children (aged 4-10 years). The database contains spontaneous narrative speech, which was obtained through storytelling tasks. All participating children told the stories three times, with a one-year interval. For the current project, only the recordings of children with DLD were used. This dataset consists of approximately 4.8 hours of speech.

For the ASR experiments, the DLD subset was divided into a training set, a test set and a validation set. The training set served two purposes: (1) as source data for the VC pipeline and (2) as training data for ASR fine-tuning. The validation set was used during the fine-tuning of the ASR model. In the final evaluation, the fine-tuned ASR system was evaluated on the test set, with performance measured using WER. The splits were constructed to preserve a balanced gender distribution, while also ensuring that the male-to-female ratio across all subsets was as consistent as possible.

Table 1: Zwitserlood Data Splits

Average Age	Train	Test	Validation
7	16	5	2
8	5	1	1
9	19	5	3
10	2	1	0
Total	42 (10 female, 32 male)	12 (3 female, 9 male)	6 (2 female, 4 male)

2.1.2 CHILDES OCSC

As target dataset for the VC, the Ohio Child Speech Corpus (OCSC) from CHILDES was used. This corpus contains speech recordings of typically developing English-speaking children from the United States. The dataset consists of 57.2 hours of speech from 303 children (164 females and 139 males, aged 4-9 years). This age range overlaps largely with the

age range of the Zwitterlood dataset. This overlap is beneficial, as acoustic and linguistic characteristics of child speech change rapidly as a function of age [11], resulting in more similar acoustic properties when age ranges are closely matched. Additionally, the relatively large size of the corpus contributes to generating diverse speech samples. Finally, the high recording quality of the dataset makes it well-suited for extracting speaker characteristics. Together, these properties make the OCSC dataset a suitable target dataset for VC.

2.1.3 Data preprocessing

Before the above mentioned datasets could be used, both the Zwitterlood dataset and the OCSC dataset had to be preprocessed. In both datasets there were recordings that contained adult speech. For the source data, this was undesirable as it could lead to the situation where the content of the source speech does not reflect the characteristics of speech from Dutch children with DLD, which would result in synthetic data with reduced child- and DLD-likeness. In the OCSC dataset it was problematic, as adult speech can distort VC; the acoustic characteristics of the adult speech can become intertwined with those of the child, leading to an impure representation of child speech. Therefore, adult speech was removed in both datasets, so that only children’s speech remained. This was achieved by segmenting the audio files into utterances with the help of the transcriptions.

The second issue in the Zwitterlood data was that it contained utterances longer than 30 seconds. Using these source utterances directly would result in synthetic utterances of similar lengths. This was undesirable, as long utterances can lead to problems in later steps of ASR fine-tuning. The utterances were therefore split into shorter fragments. The segmentation was done between sentences, so as to preserve linguistic coherence. This ensured that the resulting segments maintained a natural linguistic structure of the speech, while keeping audio lengths within the operational limits of the ASR model.

2.2 ASR model

The data created was then used to fine-tune the Whisper ASR Model from OpenAI [12]. Whisper is a large-scale pre-trained ASR system trained on 680,000 hours of multilingual and multitask speech data. This extensive and diverse training corpus makes the model very robust to variations in accents and noise. Therefore, this model is highly suitable to use as an ASR model for this study, which also concerns atypical speech of children with DLD as well as recordings containing background noise.

2.3 Experimental setup

Several ASR fine-tuning configurations were evaluated in order to investigate the impact of data augmentation through VC on ASR performance measured in terms of WER. An overview of all experimental configurations is provided in Table 2.

As a baseline, the original, non-fine-tuned Whisper model (Z0) was used. Subsequently, the ASR model was fine-tuned exclusively on the original training data from the Zwitterlood dataset to assess the effect of fine-tuning on real speech only.

The second set of experiments investigated the extent to which source-target speaker similarity influences the quality of the generated speech. Following the source-target pairing procedure described in 2.4.2, three pairing strategies were considered: most similar (SIM), random (RND), and least similar (DIS). For each strategy, the ASR model was fine-tuned on the corresponding synthetic dataset, resulting in three separate configurations. In addition,

separate experiments were conducted in which the ASR model was fine-tuned on the synthetic data in combination with the original Zwitterlood training data. Previous work has shown that, when only limited amounts of task-specific speech are available, combining original and synthetic speech can improve model generalization [13]. These experiments therefore investigated whether this complementary fine-tuning configuration can also improve ASR performance for Dutch children with DLD.

The third set of experiments examined the effect of augmentation quantity. Zhang et al. [5] reported that augmentation quantity affects child speech recognition performance and found that pairing source speakers with the most similar target speakers resulted in lower or equal WERs, compared to random and least similar pairing strategies. Based on these findings, only the most similar pairing strategy was used for the augmentation-factor experiments. Zhang et. al also reported that a two-fold augmentation factor yielded strong performance gains for child speech recognition, while larger augmentation factors did not consistently provide additional improvements. Therefore this work evaluated 0.5-fold and 1-fold augmentation factors in addition to the 2-fold setting, to determine whether comparable improvements can be achieved with smaller amounts of synthetic speech. For each augmentation factor, the ASR model was fine-tuned both on the synthetic data alone (SIM-0.5 and SIM-1) and on a combination of synthetic and original speech data (ORG+SIM-0.5 and ORG+SIM-1).

All experiments were conducted using the SpeechBrain toolkit [14]. First Whisper-large-V3 was fine-tuned on the training data with Low-Rank Adaptation (LoRA) [15] for 50 epochs. The model achieving the lowest WER on the validation set was selected as the final model. The validation set was used for this, as it consisted of unseen data and therefore prevented the risk of overfitting. A beam size of 10 was used for decoding. All ASR experiments were conducted on an NVIDIA A40 GPU.

Table 2: Experimental configurations for ASR fine-tuning.

Model ID	Data type	Pairing strategy	Augmentation factor
Z0	None	-	-
ORG	Original	-	1
SIM-2	Synthetic	Most similar	2
ORG+SIM-2	Original + synthetic	Most similar	2
DIS-2	Synthetic	Least similar	2
ORG+DIS-2	Original + synthetic	Least similar	2
RND-2	Synthetic	Random	2
ORG+RND-2	Original + synthetic	Random	2
SIM-0.5	Synthetic	Most similar	0.5
ORG+SIM-0.5	Original + synthetic	Most similar	0.5
SIM-1	Synthetic	Most similar	1
ORG+SIM-1	Original + synthetic	Most similar	1

2.4 Data generation pipeline

2.4.1 Speaker embeddings

In order to calculate the speaker similarities between source and target speakers, a key requirement was obtaining reliable and compact representations of speaker identity. Speaker

embeddings are required for this purpose, since similarity cannot be directly computed on raw variable-length speech segments [16].

To address this, speaker embeddings were extracted using the ECAPA-TDNN (Emphasized Channel Attention in Time Delay Neural Network based speaker verification) model [17]. ECAPA-TDNN is an enhanced TDNN-based speaker embedding model that builds upon the x-vector architecture and improves its performance. ECAPA-TDNN improves the embedding quality through channel attention, because of which the model can focus on informative feature channels. Because of these improvements, ECAPA-TDNN creates stronger and more robust speaker embeddings than other TDNN-based models [17].

The model was applied to all utterances of each target speaker to obtain multiple embeddings per speaker. Subsequently, the embeddings were averaged per speaker to obtain a single representative speaker embedding. This was done by computing the mean per dimension across all utterances of a speaker. Finally, in order to ensure a consistent embedding scale, the resulting averaged vector was L2-normalized.

2.4.2 Source-target speaker pairing

Another challenge in the pipeline was how to use the created speaker embeddings to do the pairing according to the three strategies. To address this, the cosine similarities between each source and target speaker were calculated based on the ECAPA-TDNN embeddings. These embeddings represent the identity of a speaker in a compact vector space, where cosine similarity is a measure of acoustic closeness between two speakers; two embeddings of similar speakers yield high cosine similarity values, while embeddings of dissimilar speakers have a lower similarity.

This was the basis of three different pairing strategies used in this research. For the 2-fold augmentation setting pairing was defined as follows. In the most similar strategy, for each source speaker the two target speakers with the highest cosine similarity were selected. In the random pairing strategy, two target speakers were selected randomly per source speaker to maximize diversity. In the most dissimilar speaker pairing strategy, for each source speaker, the two target speakers with the lowest cosine similarity were selected. This intentionally forced large speaker differences between the source and target speaker.

For 1-fold augmentation, only the single target speaker with the highest cosine similarity was selected per source speaker. For the 0.5-fold augmentation setting, half of the speakers were randomly sampled, and from this half, the speaker with the highest cosine similarity was selected for each source speaker.

2.4.3 Selecting best quality target utterances

For the VC model, only a single source and a single target utterance were required per conversion. Therefore, all source utterances were mapped to a single target utterance per speaker. The approach of concatenating all target utterances of a speaker to use as target was discarded, as this disrupted the natural speech flow. Instead, a single utterance per target speaker was selected.

Prior work has shown that degraded input speech can negatively affect VC quality [18]. Based on this, the selection of target utterances for the VC pipeline was based on the quality of the utterances. Per target speaker, only the utterance with the highest predicted perceptual quality was selected.

To estimate this quality, MBNet [19] was used to evaluate all utterances for each speaker. MBNet is a neural MOS predictor that uses a mean-bias architecture where a MeanNet

predicts the general quality and a BiasNet systematically learns to assess differences between predictors. This makes it so that MBNet scores correlate strongly to human assigned speech ratings [19].

2.5 Voice conversion

For the VC, the MeanVC-model [10] was chosen due to its computational efficiency and empirical performance. MeanVC is a relatively lightweight model, which makes it practically usable in the limited span of this ten-week research project. However, it still retains strong results in the field of speech quality, speaker similarity and efficiency [10]. Ma et al. even present MeanVC as better than other existing streaming VC models, while at the same time needing less computation power. Thus MeanVC has the right balance between implementation ease and performance for this project.

MeanVC is a zero-shot VC model that is able to extract the speech characteristics of a target speaker without any additional training on that specific speaker. The resulting synthetic speech contains the linguistic content of the source, with the acoustic characteristics of the target speaker [10].

In this work, the CHILDES Zwitterlood dataset was used as source data, while the target speech was from the CHILDES OCSC dataset. The specific speakers were selected from the created source-target speaker pairs.

2.6 Evaluation metrics

2.6.1 WER

The fine-tuned Whisper model was used on the test set and the WER was used as the evaluation metric of the ASR performance. The WER is a measure of ASR accuracy, quantifying the percentage of substitutions, deletions and insertions compared to the reference transcription [20].

The WER was chosen for multiple reasons. First, it reflects the level of speech recognition directly. This aligns with the goal of accurately recognizing speech utterances of children with DLD. Second, the WER allowed us to compare to prior work directly, as the WER is a widely used evaluation metric.

2.6.2 Subjective MOS

To evaluate perceptual speech quality, a subjective listening test based on MOS was conducted. MOS is a subjective evaluation metric where human listeners rate speech quality on a discrete five-point scale: 5 – Excellent, 4 – Good, 3 – Fair, 2 – Poor, 1 – Bad. The MOS is computed as the arithmetic mean over all listener ratings [21]:

$$\text{MOS}(u) = \frac{1}{K} \sum_{k=1}^K s_{u,k}$$

Additionally, for each MOS value, a 95% confidence interval was calculated as follows:

$$CI = \bar{X} \pm t_{\alpha/2, N-1} \cdot \frac{S}{\sqrt{N}}$$

Subjective MOS evaluation was included as it remains one of the most widely used and accepted methods for assessing perceived speech quality in speech processing research [21].

Moreover, objective signal-based metrics often correlate poorly with human perceptions of speech quality [22]. MOS predicting models, such as AMOS are promising, but these systems are typically trained on large amounts of typical adult speech and may therefore give less reliable ratings to non-standard speech, such as child speech with DLD. By relying on human listeners, subjective MOS evaluation mitigates this risk and provides a direct assessment of the perceived quality of the source speech and the synthetic speech.

The listening test included nine participants (five female and four male), aged between 18 and 60 years. All participants were native Dutch speakers. Prior to the experiment, participants completed a short questionnaire collecting demographic information and information regarding hearing status. Responses indicating hearing issues were used as an exclusion criterion to ensure data quality.

For the evaluation, ten source utterances were randomly selected from the Zwitterlood dataset. For each of these utterances, three synthetic speech utterances which were created using VC with the three speaker-target pairing strategies were added to the listening test. Specifically, the following four stimuli were presented per source utterance:

- The original source utterance
- A VC-generated utterance using the most similar strategy
- A VC-generated utterance using a randomly selected strategy
- A VC-generated utterance using the least similar strategy

This resulted in a total of 40 stimuli. Each participant evaluated the same 40 stimuli. Each set of four stimuli corresponding to a single source utterance was presented on a single page. The order of the four stimuli within each page was randomized per set to reduce ordering effects and listener bias.

Participants were asked to evaluate each stimulus using the five-scale MOS ratings on three dimensions: naturalness, intelligibility and childlikeness, which were defined as follows: **Naturalness:** how natural the speech sounds.

Intelligibility: how easy it is to understand what the speaker is saying.

Childlikeness: how much the voice sounds like that of a child.

The order of these rating dimensions was kept constant across all stimuli (first naturalness, then intelligibility and finally childlikeness). Each participant rated all stimuli, resulting in a within-subjects experimental design.

2.6.3 AMOS

In order to assess the quality of speech generated through VC, AMOS values were computed using MBNet [19], a neural network model trained to predict the human MOS ratings for speech utterances. For each source-target pairing strategy, MBNet was used to predict MOS values for all synthesized utterances individually. Then, the predicted scores were averaged per strategy by computing the arithmetic mean over all utterances generated with the respective strategy.

AMOS was selected as an evaluation metric, as it provides full coverage of the generated dataset, in contrast to the subjective MOS evaluation, which only covered a subset of the synthesized data. This complete data coverage reduces the risk of sampling bias and yields a more representative estimate of the overall performance of the system. Moreover, AMOS enabled scalable and efficient evaluation of perceptual speech quality without requiring human participants, making it a practical complement to the subjective MOS.

3 Results

3.1 Performance of the ASR model

Table 3 presents the WER results of all fine-tuning configurations. Overall, fine-tuning on the original speech data (ORG) yielded the best performance, reducing the WER by 14.9 percentage points compared to the non-fine-tuned baseline (Z0). This suggests that real speech from children with DLD remains the most effective fine-tuning data for the ASR model.

When combining original and synthetic speech, the ASR performance generally improved compared to fine-tuning on synthetic speech exclusively. Notably, ORG+DIS-2 outperformed DIS-2 by 16.7 percentage points. These results indicate that using synthetic speech is most beneficial as supplementary fine-tuning data, instead of a replacement for original speech.

Finally, the quantity experiments showed that decreasing the amount of synthetic data used during fine-tuning can lead to improved performance. In the configurations fine-tuning on original and synthetic data, the WER decreased by 10.3 percentage points, when reducing the augmentation factor from 2 to 0.5. This suggests that maintaining a smaller partition of synthetic speech relative to real speech may be more beneficial for fine-tuning an ASR model.

Table 3: WERs per ASR fine-tuning configuration. Bold indicates the lowest WER.

Model	Data type	Pairing	Aug.	Hours	WER%
Z0	None	-	-	-	46.5
ORG	Original	-	1	3.4	31.6
SIM-2	Synthetic	Most similar	2	6.8	49.5
ORG+SIM-2	Original + synthetic	Most similar	2	10.2	49.5
DIS-2	Synthetic	Least similar	2	6.9	50.3
ORG+DIS-2	Original + synthetic	Least similar	2	10.3	33.6
RND-2	Synthetic	Random	2	6.4	45.0
ORG+RND-2	Original + synthetic	Random	2	9.8	39.1
SIM-0.5	Synthetic	Most similar	0.5	1.4	48.9
ORG+SIM-0.5	Original + synthetic	Most similar	0.5	4.8	39.2
SIM-1	Synthetic	Most similar	1	3.4	52.0
ORG+SIM-1	Original + synthetic	Most similar	1	6.8	40.3

3.2 Perceptual quality through subjective MOS

The source speech received substantially higher ratings than the synthetic speech across all three quality dimensions. The average difference between the ratings of the source and VC conditions was approximately 1.95 MOS points, indicating a clear perceptual degradation introduced by VC. This suggests that the synthesized speech is perceived as less intelligible, less childlike and less natural than the original Zwitsersloot recordings.

When considering the VC conditions alone, childlikeness received the highest mean MOS rating (2.33), followed by intelligibility (1.94) and naturalness (1.90). This pattern is reflected in the source condition, where childlikeness was also rated higher than the other

perceptual dimensions. However, the gap between source and synthetic speech was smallest for childlikeness (approximately 1.84 MOS points), compared to intelligibility and naturalness, which had an average difference of 1.98 and 2.00 respectively. This suggests that, while overall quality decreases substantially after VC, childlikeness is preserved better than intelligibility and naturalness.

Table 4: Mean ratings (1–5 scale) with 95% confidence intervals for intelligibility, childlikeness, and naturalness across conditions.

Condition	Intelligibility	Childlikeness	Naturalness
Source	3.92 ± 0.22	4.17 ± 0.19	3.90 ± 0.19
Most Similar	1.93 ± 0.17	2.31 ± 0.19	1.85 ± 0.15
Random	1.92 ± 0.19	2.34 ± 0.24	1.93 ± 0.19
Least Similar	1.96 ± 0.18	2.34 ± 0.22	1.91 ± 0.16

3.3 Perceptual quality through automatic MOS

Figure 1 depicts the distribution of AMOS-predicted scores across the source speech and the speech generated by VC with the use of source-target pairing strategies. Overall, there was a clear separation between the source audio and the VC-generated speech. The source condition exhibited a relatively narrow distribution and received the lowest AMOS values. The minimal distance between the mean score of the source speech and the mean scores of the synthetic speech was 0.83. This suggests that the original recordings are of relatively low perceptual quality and that there is limited variability in this quality across utterances. In contrast, the model predicted substantially higher MOS values for all VC conditions, suggesting that this speech has better perceived quality.

3.4 The effect of source-target speaker similarity

3.4.1 Performance of the ASR model

When fine-tuning on synthetic data exclusively, the differences between the source-target pairing strategies were relatively small, as can be seen in table 3. The WERs for the most similar (SIM-2), random (RND-2) and dissimilar (DIS-2) strategies were all comparable. This indicates that the choice of speaker pairing strategy has limited impact when fine-tuning on synthetic speech only.

In contrast, when fine-tuning on a combination of original and synthetic speech, larger differences between strategies emerged. Interestingly, the least similar speaker pairing strategy (ORG+DIS-2) performed best, outperforming both random (ORG+RND-2) and most similar (ORG+SIM-2) models. ORG+SIM-2 performed worst and showed no improvement compared to Z0, suggesting that fine-tuning using the most similar strategy is less effective.

3.4.2 Subjective MOS

The three source-target pairing strategies produced highly comparable MOS scores, as shown in table 4. Differences were minimal across all perceived quality metrics. The largest observed difference was 0.08 for naturalness between the most similar and random conditions.

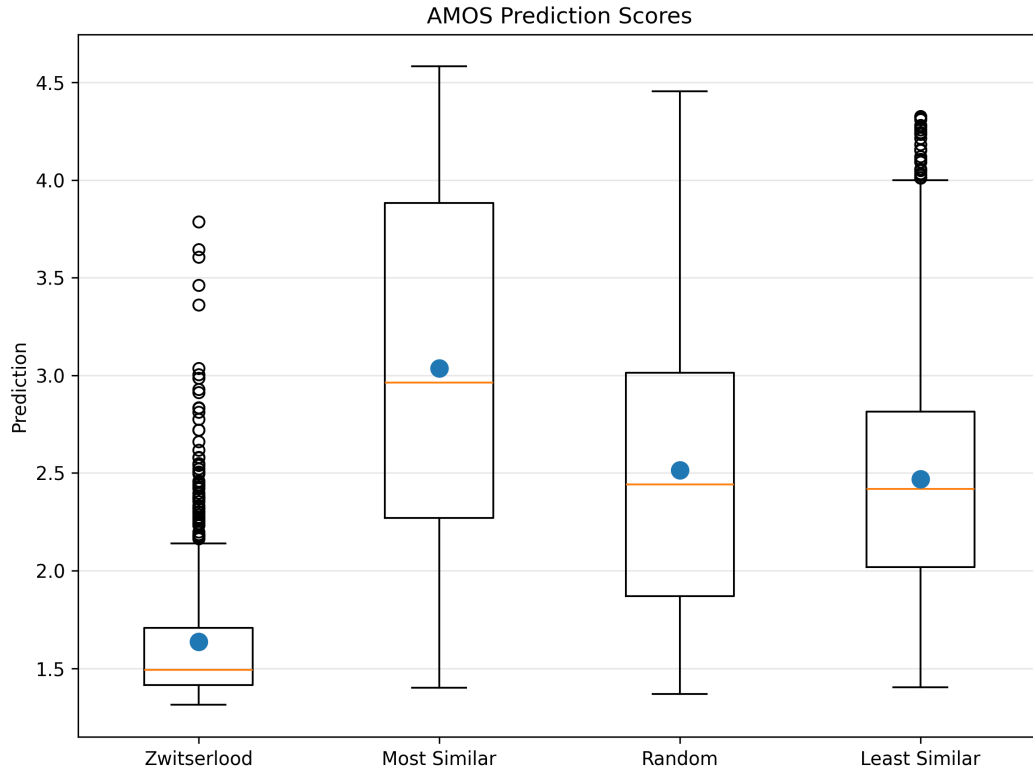


Figure 1: AMOS prediction scores across source-target speaker conditions. The blue filled markers indicate the mean MOS per condition.

This suggests that the choice of speaker pairing strategy has limited influence on the perceived quality of synthesized speech.

3.4.3 AMOS

In figure 1 it can be seen that among the VC strategies, the most similar source-target pairing received the highest MOS predictions and showed the widest distribution. This suggests that the most similar pairing strategy has the best perceived quality, albeit with increased variability in the utterances.

Random and least similar pairing strategies received lower MOS predictions, with distributions highly comparable to each other. This indicates that there is no systematic advantage of using one strategy over the other in terms of perceived quality.

4 Responsible Research

Most ASR systems are trained on typical adult speech, making them perform less well for children with DLD. The development of ASR systems that are accessible to diverse user groups is an important ethical responsibility. Inclusive speech technologies can help

this underrepresented group of children by lowering communication barriers, creating more learning opportunities and improving accessibility of technologies.

The research is largely reproducible, but not fully end-to-end deterministic. Most components are open and can be replicated under the same experimental setup. The datasets used are publicly available. Using the same preprocessing steps ensures that the input to the experiments is consistent. Additionally, all technologies used in the research are publicly available. However, the human evaluation using MOS is not completely reproducible, as subjective assessments can differ between different participants and repeated experiments. This reduces strict reproducibility.

Ethical considerations were taken into account, particularly for the human evaluation study. All participants provided informed consent prior to participation via a consent form in accordance with ethical research guidelines. Participant data was handled confidentially and anonymized where applicable to ensure privacy and compliance with ethical standards.

In line with the GenAI guidelines for end projects from the TU Delft, this section discloses the use of generative AI tools. AI was exclusively used as a supportive tool, not as binding decision-maker in the research or analysis process. Specifically, ChatGPT [23] was used to assist in coding as well as in writing this paper.

The trustworthiness of AI-generated output was assessed critically, with particular attention to possible biases and limitations of the model used. All generated output was carefully reviewed and verified before being applied.

In the coding process, AI was used to assist in debugging, to provide suggestions and to propose ideas on how to solve a problem. In the writing process, AI was used in the preparation stage for generating structural ideas and for suggesting relevant paragraph content. Moreover, it was used to provide feedback on draft sections and to give suggestions to improve clarity and writing style. Finally, it was used to filter out typographical, spelling, and grammatical errors. No unreleased, personal, or sensitive data was entered into the AI-tool.

5 Discussion

5.1 Impact of synthetic speech augmentation on ASR performance

This study examined to what extent cross-lingual VC can improve the performance of ASR for Dutch children with DLD, compared to a baseline without data augmentation. Previous research has shown that synthetic speech can improve ASR performance substantially. However, in this study, fine-tuning the ASR model with synthetic data did not lead to better performance than fine-tuning it exclusively on real speech from Dutch children with DLD.

Several factors may explain the lack of improvement in ASR performance. A first explanation is the presence of artifacts introduced by VC, including robotic sounding segments and speech distortions. These artifacts may have reduced the naturalness and overall quality of the generated speech. This interpretation is supported by the MOS results, which showed that naturalness, intelligibility, and childlikeness were all rated lower for the synthesized speech compared to the original source speech.

A second explanation concerns the preprocessing of the source data. A part of the source data contained adult speech, which was removed by segmenting and concatenating the audio. This preprocessing may have introduced unnatural speech transitions, which potentially had a negative effect on VC quality.

The results across augmentation folds provide further insight into the role of the quality of the synthetic data. In experiments where the ASR model was fine-tuned on original and synthetic data, smaller augmentation folds performed better than larger augmentation folds. This might be explained by the fact that in smaller augmentation folds, less low-quality or unnatural speech is used during fine-tuning. This is supported by the observation that, in some experimental configurations, the ASR model fine-tuned on larger amounts of synthetic data increased the WER, performing worse even than the out-of-the-box baseline.

5.2 Subjective MOS

The MOS evaluation did not show significant differences in perceived naturalness between the synthetic data of the three source-target speaker strategies. One possible explanation for this is the limited statistical power of the listening test due to the relatively small number of participants and the fact that the test only covered a subset of the original and synthesized data. As a result, the absence of significant differences should be interpreted with caution; larger scale evaluations might capture distinctions that this study was not able to.

5.3 AMOS

Notably, the source data received substantially lower AMOS predictions than the synthetic data. This is unexpected, as real speech is generally found to sound more natural than synthetic speech. Moreover, this finding contradicts the results of the subjective MOS, where the source speech performed significantly better compared to the synthesized data.

A possible explanation for the low AMOS scores of the source data lies in the cuts and concatenations present in the source data itself. Because MBNet is trained with frame-level supervision [19], it may be sensitive to local acoustic patterns. As a result, such abrupt transitions may be penalized more strongly by the model than by human participants in the MOS evaluation. During VC, these irregularities might have been smoothed out, as the Mel-spectrograms generated by flow matching models that are used in MeanVC suffer from over-smoothing [10]. This smoothing might have suppressed the abrupt transitions, leading to higher AMOS scores for the synthesized speech than for the original source speech.

Furthermore, the cause of this contradictory finding can lie in the fact that MOS estimators have several known limitations. For instance, they are trained primarily on typical adult speech and may not generalize well to atypical or disordered speech. As a result, the results of AMOS should be interpreted cautiously.

5.4 Influence of speaker pairing

While MOS did not reveal significant differences between the three source-target speaker strategies in VC, differences did emerge in both AMOS and ASR performance. In the AMOS evaluation, the mean predicted MOS of the data created using the most similar speaker strategy was higher than that of the other two strategies, which showed comparable performance. This finding is in line with the work of Zhang et al. [5].

However, in contrast to the findings of Zhang et al., the least similar source-target speaker pairing resulted in the best ASR performance, while the most similar speaker pairing performed worst. One possible explanation for this is that more speaker diversity, which might be created by the dissimilar pairing strategy, may expose the ASR model to a wider range of acoustic variations, leading to more robustness. This effect may be especially beneficial for speech from children with DLD, which in itself contains much variability.

5.5 Discrepancy between WER, MOS and AMOS

The results show a mismatch between perceptual quality metrics and ASR performance. The least similar speaker pairing achieved the lowest WER overall, while also receiving the lowest AMOS scores of all strategies and comparable values to the other strategies in MOS. In contrast, the most similar speaker pairing scored the highest in AMOS, but had the worst ASR performance.

These findings suggest that perceptual quality and usefulness for ASR training are not necessarily aligned. A possible explanation for this is that speech that is perceived as qualitatively better does not necessarily contain the acoustic variability that benefits ASR fine-tuning. A second explanation is the limited statistical significance of the MOS evaluation. Finally, an explanation lies in the limitations of the AMOS model, MBNet [19]. MBNet is trained primarily on typical adult speech and might therefore not be a good indicator for speech of children with DLD.

5.6 Limitations

Apart from the limitations mentioned above, several additional limitations were identified. First, the study relies on a single ASR architecture and a single VC model, making the results dependent on these specific architectures. Since the VC model has a large influence on the generated speech, future work should include an alternative VC model to compare the results to our findings, or use of multiple VC systems to avoid dependency on a single model. Similarly, using additional ASR architectures would improve the generalizability of the results.

Another limitation is the inconsistencies of the OCSC speech corpus. In the OCSC dataset, transcriptions did not always fully align, leading to cases where adult speech segments were inadvertently included in the data. This contamination may have affected the voice conversion process, particularly in preserving child-specific vocal characteristics.

6 Conclusions and Future Work

This study on ASR performance improvement through VC showed that fine-tuning the ASR model with synthetic speech did not lead to better performance than fine-tuning it exclusively on real speech from Dutch children with DLD. The perceptual evaluation metrics revealed contrasting outcomes. Subjective MOS ratings for VC-generated speech were considerably lower than for the source speech, while AMOS assigned higher quality scores to the synthetic speech than to the source recordings. With regard to source-target speaker similarity, the findings suggest that its influence on the perceptual quality is relatively limited. However, its effect on ASR performance was more pronounced. In the evaluation of fine-tuning on a combination of original and synthetic speech, the least similar pairing strategy achieved the best performance.

Future work should explore a wider range of ASR and VC models to reduce model-specific bias and improve robustness of these findings. Extending this approach to additional languages could also be beneficial, as this would allow the use of more datasets and enable cross-lingual or multilingual voice conversion studies. Finally, further investigation is needed to understand why dissimilar speaker pairings outperformed similar ones in ASR performance, as this finding challenges common assumptions in VC-based data augmentation.

7 Acknowledgements

The author would like to thank Yuanyuan Zhang for her contribution to this work. While the author designed the experimental setup, defined the experiments, and prepared the data for fine-tuning, Yuanyuan Zhang carried out the implementation of the ASR fine-tuning experiments and calculated the corresponding WER results. Her support in executing these experiments was essential for the completion of this study.

References

- [1] S. Shah Nawazuddin et al. Developing children's asr system under low-resource conditions using end-to-end architecture. *Digital Signal Processing*, 146:104385, 2024.
- [2] X. Wan. Improving the performance of automatic speech recognition for children with developmental language disorders. Master's thesis, Delft University of Technology, 2025.
- [3] Purdue University Child Language Research Projects. Developmental language disorder (dld) in the spotlight. Newsletter, 2023. Language Links Newsletter, Spring 2023, Department of Speech, Language, and Hearing Sciences.
- [4] Isabelle Rapin. Practitioner review: Developmental language disorders: A clinical update. *Journal of Child Psychology and Psychiatry*, 37(6):643–655, 1996.
- [5] 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.
- [6] Takuhiro Kaneko and Hirokazu Kameoka. Parallel-data-free voice conversion using cycle-consistent adversarial networks. *arXiv preprint arXiv:1711.11293*, 2017.
- [7] Yunje Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*, 2017.
- [8] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion. *CoRR*, abs/1907.12279, 2019.
- [9] Yi Zhao, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, and Tomoki Toda. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion, 2020.
- [10] Guobin Ma, Jixun Yao, Ziqian Ning, Yuepeng Jiang, Lingxin Xiong, Lei Xie, and Pengcheng Zhu. Meanvc: Lightweight and streaming zero-shot voice conversion via mean flows, 2025.
- [11] Alexandros Potamianos, Shrikanth Narayanan, and Sungbok Lee. Automatic speech recognition for children. In *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, pages 2371–2374, 1997.
- [12] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pages 28492–28518. PMLR, 2023.

- [13] S. Shah Nawazuddin, Nagaraj Adiga, Kunal Kumar, Aayushi Poddar, and Waquar Ahmad. Voice Conversion Based Data Augmentation to Improve Children's Speech Recognition in Limited Data Scenario. In *Interspeech 2020*, pages 4382–4386, 2020.
- [14] SpeechBrain Contributors. Speechbrain on hugging face. <https://huggingface.co/speechbrain>, 2026. Accessed: 2026-06-10.
- [15] Zhesu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. Lora-whisper: Parameter-efficient and extensible multilingual asr, 2024. arXiv preprint.
- [16] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, 2018.
- [17] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, pages 3830–3834, 2020.
- [18] Xining Song, Zhihua Wei, Rui Wang, Haixiao Hu, Yanxiang Chen, and Meng Han. Degrading voice: A comprehensive overview of robust voice conversion through input manipulation, 2025.
- [19] Yichong Leng, Xu Tan, Sheng Zhao, Frank K. Soong, Xiangyang Li, and Tao Qin. Mbnet: MOS prediction for synthesized speech with mean-bias network. *CoRR*, abs/2103.00110, 2021.
- [20] Korbinian Kuhn et al. Measuring the accuracy of automatic speech recognition solutions, 2024.
- [21] International Telecommunication Union. Recommendation itu-t p.800: Methods for subjective determination of transmission quality. Technical Report P.800, ITU-T, August 1996. Previously CCITT Recommendation, approved 30 August 1996, Study Group 12.
- [22] A.W. Rix. Perceptual speech quality assessment - a review. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages iii–1056, 2004.
- [23] OpenAI. Chatgpt. <https://chat.openai.com>, 2026. Large language model used for code assistance and writing support.