Improving the Performance of Automatic Speech Recognition for Children with Developmental Language Disorders

Master thesis report

Xin Wan

Improving the Performance of Automatic Speech Recognition for Children with Developmental Language Disorders

by

Xin Wan

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday, November 3, 2025, at 11:00 AM.

Student number:	5633214	
Project duration:	March 25, 2024 – November 3, 2025	
Thesis committee:	Dr. O. E. Scharenborg,	TU Delft, thesis advisor
	Dr. J. Sun,	TU Delft, external committee supervisor
	Dr. T. Viering,	TU Delft, external committee supervisor

Contents

1 Introduction

1.1. Motivation

Automatic Speech Recognition (ASR) has become a significant technique in improving human-machine interaction, playing a critical role in accessibility [1], communication [2], education [3], and entertainment [4]. Most of the research in ASR is aimed primarily at the adult population, and reported ASR systems have achieved high rates of recognition of adult speech [5]. However, ASR systems that are trained using data from adults see a significant decrease in performance when recognizing speech from children [6]. The acoustics and linguistic characteristics, spectral and temporal factors, of adults and children exhibit notable differences [7, 8]. Variation in these characteristics leads to a mismatch between the speech patterns of children and adults [5]. The main reason for these variations is the morphological and anatomical disparities in the vocal tract, together with the restricted control that children have over prosodic features such as pitch, intensity, speed, and intonation [5]. The disparity is more pronounced in children with speech difficulties, as their speech patterns may markedly differ from those of generally developing peers [9, 10]. Currently, popular state-of-the-art commercial ASR systems (Amazon Web Services Transcribe, Google Cloud Speech, and IBM Watson Speech-to-Text) show low ASR accuracy when processing impaired speech [11]. These systems are mostly trained on huge amounts of typical speech, which means that they do not work accurately for impaired speech. Despite some recent developments in personalization of ASR models [12], data augmentation and transfer learning approaches[13], and the creation of specialized impaired-speech datasets such as UASpeech [14], the accuracy of ASR systems for impaired speech remains much lower compared to typical speech [15].

Developmental Language Disorder (DLD), a prevalent condition affecting around 5–8% of preschool-aged children [16]. It encompasses many disorders in which a child has notable deficits in speech acquisition, comprehension, or language utilization, all crucial for effective communication and academic achievement. Children with DLD may demonstrate several linguistic deficiencies, including difficulties in producing speech sounds accurately, using vocabulary effectively, and constructing grammatically correct sentences [17, 18]. These issues are manifest in their speech patterns, which may markedly diverge from those of their typically developing peers.

Although ASR technology has the potential to improve communication and learning for children with DLD, its implementation is limited by two primary limitations. First, there exists a deficiency of publicly accessible, high-quality speech datasets from children with DLD, mostly attributable to ethical and privacy issues related to the collection and distribution of data from vulnerable groups [19, 20, 21]. This makes DLD speech a low resource. Second, there is a notable mismatch between the speech of children with DLD and the speech on which current ASR systems are trained—typically large-scale datasets of typcial adult speech.

Data augmentation [22] is a technique used in machine learning to increase the amount of data available for training models without actually collecting new data [23, 24], which has been demonstrated to be effective in addressing the issue of data sparsity [25, 26]. Recently, some research has shown that using data augmentation such as speed perturbation (SP) [27, 28] and vocal tract length perturbation (VTLP) [29, 30] increases the performance for both child and impaired speech recognition.

Moreover, transfer learning approaches, such as fine-tuning [31], aim to solve these problems by initially training a model on a larger corpus that is not specific to the target domain. The learned features are then adapted and used to train a network on a smaller dataset that is specific to the target domain [32]. Previous research has demonstrated the efficacy of these techniques for languages with limited resources [33, 34, 35], non-native accent speech [36, 19] and disorder speech of adults [37].

1.2. Research Questions

This study aims to improve the ASR performance for the speech of children with DLD using data augmentation and transfer learning, in contrast to previous research that focused primarily on typical adult speech. It also examines performance on typical child speech to ensure that improvements in DLD child speech do not come at the expense of recognition accuracy in typical child speech. It evaluates whether these techniques improve recognition of DLD child speech without negatively impacting performance on typical child speech. Achieving this balance is essential for developing robust and generalizable systems suitable for practical educational and clinical settings, where both types of speech may be present.

The main research question is

- How can ASR performance be improved for the speech of children with DLD while maintaining recognition accuracy on typical child speech?

We decompose this question into smaller research questions:

- **RQ1**: To what extent do data augmentation techniques, such as SP and VTLP, improve ASR performance for the speech of children with DLD without degrading recognition accuracy on typical child speech?
- **RQ2**: To what extent does transfer learning through fine-tuning improve ASR performance for the speech of children with DLD while maintaining accuracy on typical child speech?
- **RQ3**: To what extent does combining data augmentation (SP and VTLP) with transfer learning (fine-tuning) improve ASR performance for the speech of children with DLD without negatively impacting recognition accuracy on typical child speech?

1.3. Outline

In this thesis, Chapter 2 provides essential background knowledge for a comprehensive understanding of this thesis. Chapter 3 explores the methods. Chapter 4 describes the experiments designed to investigate the research questions. Chapter 5 presents the experimental results. Chapter 6 provides discussions, conclusions, and future work based on the experimental results.

2 Background

This chapter presents the essential knowledge for this thesis and discusses relevant studies. I start with an introduction to developmental language disorder (DLD) in section 2.1. In Section 2.2, we provide background information about automatic speech recognition, covering traditional ASR systems (Section 2.2.1) as well as end-to-end ASR models (Section 2.2.2). The ASR-related data augmentation methods are described in Section 2.3, where I start with the speed perturbation in Section 2.3.1 and follow with vocal tract length perturbation in Section 2.3.2. Fine-tuning refers to ASR in low-resource situations and is described in detail in Section 2.4. Lastly, Section 2.5 describes the evaluation metric used in this thesis.

2.1. Developmental Language Disorder

Developmental Language Disorder [38] is a neurodevelopmental condition which hinders a child's ability to learn and use spoken language in the absence of any apparent neurological, sensory or cognitive impairment. Approximately 5–8% of children are affected by DLD and tend to continue into teenage years and adulthood [16]. The primary challenges involve the child's ability to use words and construct sentences to convey meaning, however, many children also struggle with understanding language (receptive language) [39]. DLD can influence various aspects of language, and the extent of impairment in these areas may differ from one child to another [40]. Despite efforts to classify distinct subtypes, such efforts have typically not produced clear categories [41]. The study [42] advised that specific language impairments should be evaluated and recorded for each individual child, acknowledging that children can exhibit diverse combinations of challenges [39]. The potential areas impacted include:

- **Phonology** – Younger DLD children tend to make speech sound errors, they may struggle to differentiate between specific speech sounds, resulting in the production of 'cake' as 'tate' [43]. Phonological difficulties that have persisted over time will result in unclear speech and affect children's earliest experiences of learning to read [44, 45].
- **Grammar** – Children with DLD have trouble constructing grammatically correct sentences (syntax) and using word structure to give meaning (morphology), and make errors such as missing tense markers or wrong word order (e.g.,'me jump here' instead of 'I jumped here'). This difficulty also affects comprehension, especially when faced with complex sentence structures or grammatical markers [46, 47, 48].
- **Semantics** – Vocabulary development is delayed. Children may use vague or overgeneralized words and, one word might have a lot of meanings (eg. "cold" temperature, sickness or emotional state) [49, 50].
- **Word Finding** – Some children find it difficult to retrieve known words during speaking. This "tip-of-the-tongue" phenomenon can interfere with fluency and narrative communication skills, reducing expressive vocabulary [51, 52].
- **Pragmatics** – DLD can also affect the use of language in social situations. Children may have difficulty staying on topic, understanding figurative language, or adjusting speech according to the listener or the setting [53].
- **Discourse** – Impairments in language organization above the sentence level can lead to problems in narrative telling, producing events sequentially, or even understanding discourses as a whole [54].
- **Verbal Memory And Learning** – People with DLD typically have a lower verbal working memory capacity. Speaking and processing long strings of words is difficult, and learning new ones likewise [55, 56, 57].

2.2. Automatic Speech Recognition

ASR is a computational task that transforms a signal that contains spoken language into the corresponding written text. ASR systems analyze audio waveforms and extract relevant linguistic features through multiple processing stages. ASR has evolved considerably, transitioning from early rule- and statistical model-based systems to current ASR models with deep learning and end-to-end neural architectures [58, 59, 60, 61].

The basic idea of ASR is to find the most probable sequence of words W from acoustic feature vectors X extracted from an input audio signal. This is typically modeled as a maximum a posteriori (MAP) estimation problem as Equation 2.3:

$$\hat{W} = \arg\max_{W} P(W|X) \tag{2.1}$$

Using Bayes' theorem, this can be decomposed as Equation 2.4:

$$\hat{W} = \arg\max_{W} P(X|W)P(W) \tag{2.2}$$

Where:

- $P(X|W)$: Likelihood of the acoustic signal given a word (modeled by the acoustic model)
- $P(W)$: Prior probability of the word sequence (modeled by the language model) [58]

2.2.1. Traditional ASR Model

Figure 2.1 illustrates the architecture of the traditional ASR model, which consists of three major components, the acoustic model, the lexicon, and the language model. The acoustic model maps acoustic features to the probabilities of phonemes. Second, the lexicon provides a dictionary that maps sequences of phonemes to words. Finally, the language model assigns probabilities to sequences of words to generate coherent sentences. Although each component may be developed or trained separately, they are integrated during decoding to produce the final transcription.

While traditional ASR models, such as Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) systems [62] and later Deep Neural Networks - Hidden Markov Model (DNN-HMM) hybrid systems [59], have achieved successful recognition performance, these models suffer from major drawbacks. First, they rely heavily on handcrafted acoustic features (e.g., Mel-frequency cepstral coefficients), which limits their ability to capture complex and abstract representations of speech. Second, they require pronunciation lexicons that depend on expert linguistic knowledge and often fail to adequately capture the variability of pronunciation between speakers, dialects, and spontaneous speech. Finally, their robustness is limited when faced with variability in speech, such as different acoustic environments, background noise, accents, speaking styles, or age-related speech characteristics, making them less effective in real-world or diverse speaker conditions [59, 60, 63, 64].

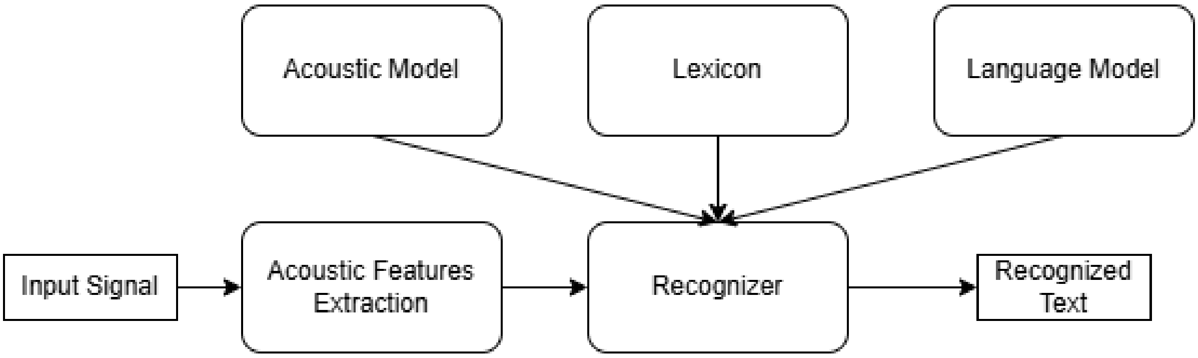

Figure 2.1: The traditional ASR architecture

2.2.2. End-to-End ASR Model

End-to-End (E2E) models were developed to address the limitations of traditional ASR systems. Unlike traditional ASR, which requires separate components for acoustic modeling, pronunciation lexicons, and language modeling, E2E approaches integrate these into a single neural network that directly maps input speech to text. This unified architecture eliminates the need for hand-crafted features and expert-designed lexicons, allowing the model to automatically learn representations and mappings from data. As a result, E2E systems are often more robust to speech variability [65, 28] and better suited for large-scale real-world applications [66, 67]. Figure 2.2 illustrates the architecture of a general E2E ASR model.

One of the greatest advantages E2E models offer is the lack of the need for lexicon dictionaries. Without using a fixed phonetic dictionary or expert linguistic input, these models can be trained end-to-end on speech data and its transcription. Thus, it helps in the training of ASR systems for low-resource languages, that is, where such resources are limited or missing.

In addition, E2E models eliminate the need for word alignments in the training process. Unlike traditional ASR systems that rely on forced alignments between acoustic frames and linguistic units, E2E approaches can be trained directly from paired audio and transcription.

Three major paradigms have gained significant popularity and form the backbone of most SOTA E2E ASR systems:

Connectionist Temporal Classification (CTC)

Although CTC is technically a loss function rather than a standalone architecture, models trained with CTC loss have become a fundamental paradigm in E2E ASR. It is capable of dealing with the alignment between input acoustic sequences and output label sequences without relying on pre-segmented data. CTC allows some flexibility by inserting a blank token and uses dynamic programming to sum

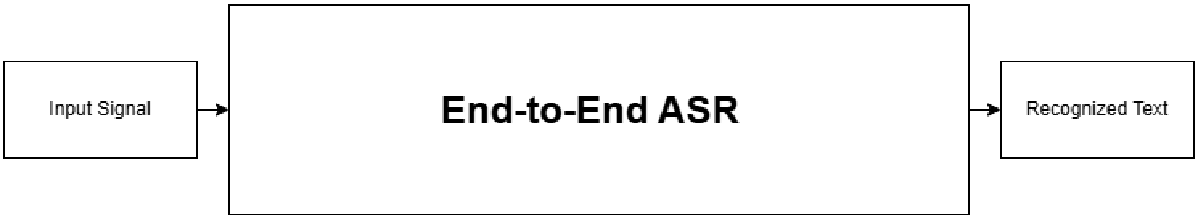

Figure 2.2: The E2E ASR architecture

the probabilities for all possible cross-alignment sequences. Although CTC is efficient and fast, it assumes conditional independence of output labels given the input, thus limiting its ability to model longer contexts across tokens [68].

Attention-Based Encoder-Decoder Models

This architecture is inspired by the translation of neural machines. It has an encoder that converts the audio signal into high-level features and a decoder that predicts the output tokens one at a time. In each decoding step, the decoder uses attention to focus on certain relevant parts of the input sequence, adding soft alignment between the input and output. These architectures can effectively model long-range dependencies and complex contextual relationships [69].

RNN-Transducer (RNN-T)

The RNN-Transducer is an extension of CTC that incorporates a prediction network that models dependencies between output tokens, while jointly modeling alignment and prediction, making it particularly useful for real-time ASR needs. Compared to CTC, RNN-T does not assume independence among output tokens, and this has demonstrated improved timestamp performance in noisy and conversational environments [70].

These three paradigms are highlighted because they represent the main approaches to end-to-end ASR, each with complementary strengths. CTC-based models provide fast and straightforward training by eliminating the need for frame-level alignments between audio and texts, making them efficient for large datasets [71]. Attention-based models aim to model long-range dependencies and contextual information, allowing them to model complex linguistic structures effectively [72]. RNN-T balances alignment modeling and token dependencies, supporting streaming recognition and low-latency applications [73].

These models have been integrated into popular toolkits and frameworks like ESPnet [74], OpenAI Whisper [75], wav2vec 2.0 [76], and DeepSpeech [77], all of which modify and enhance these architectures to address various challenges, including low-resource settings, multilingual modeling, and domain adaptation.

2.3. Data Augmentation Techniques in ASR

2.3.1. Speed Perturbation

Speed perturbation is a data augmentation technique that alters the temporal properties of a speech signal by modifying the playback speed of the audio without changing its pitch. This technique is effective in simulating natural variability in speech rate between different speakers or utterances. In ASR training, speed perturbation is typically applied using speed factors such as 0.9 and 1.1, generating slower and faster versions of the original utterance [26].

Mathematically, if $x(t)$ is the original audio signal, speed perturbation modifies the signal as Equation 2.3:

$$x'(t) = x(\alpha t) \tag{2.3}$$

where α is the speed factor. For example, $\alpha = 0.9$ slows down the audio, and $\alpha = 1.1$ speeds it up [26]. By speeding up or slowing down speech, SP creates variations in the temporal characteristics of the speech signal, which can help the ASR model become more resilient to differences in speech rate. By generating additional training examples with different speech rates, SP can help the model learn to recognize speech that is faster or slower than the original training data. This is especially important in the context of ASR for children, as children's speech often exhibits greater variability in speech rate compared to adult speech [26].

2.3.2. Vocal Tract Length Perturbation

Vocal Tract Length Perturbation is another data augmentation technique. It simulates differences in speaker vocal tract anatomy, a key factor in determining the formant frequencies of speech sounds by warping the frequency axis of the audio signal [78]. The core idea is to apply a non-linear transformation to the frequency f of the signal. A commonly used VTLP transformation is shown as Equation 2.4:

$$f' = \begin{cases} \alpha f, & 0 \leq f \leq f_0 \\ \\ \frac{f_{\text{max}} - \alpha f_0}{f_{\text{max}} - f_0}(f - f_0) + \alpha f_0, & f_0 < f \leq f_{\text{max}} \end{cases} \tag{2.4}$$

where:

- f is the original frequency,
- f' is the warped frequency,
- α is the warp factor (e.g., 0.9 to 1.1),
- f_0 is a threshold frequency (typically around 4.8 kHz),
- f_{max} is the maximum frequency (e.g., 8 kHz).

VTLP works by warping the frequency axis of the speech signal, effectively altering the perceived length of the vocal tract of the speaker. This technique can help ASR models to generalize better to different speakers by making them less sensitive to variations in the length of the vocal tract.

This warping simulates the speech characteristics of speakers with different vocal tract lengths—shorter tracts (e.g., in children) result in higher formant frequencies, while longer tracts (e.g., in adults) result in lower ones. VTLP improves model performance under speaker-diverse conditions by helping ASR systems to learn to generalize across spectral variability [78].

2.4. Finetuning in ASR

For ASR, fine-tuning can provide good benefit if the target speech data differ from the original data used to train the models. SOTA pre-trained models (e.g., based on the Transformer architectures, e.g., Whisper [79], wav2vec 2.0 [76]) are typical trained large corpora of adult speech, learning general acoustic and linguistic representations. Fine-tuning adapts these representations to better reflect the target speech. This is usually done by continuing training on the target dataset, with strategies depending on the task and the amount of available data. If the pre-trained model already includes substantial examples of the target speech data, fine-tuning may provide only limited improvements and should be done cautiously to avoid overfitting [80].

- **Full fine-tuning**: Update all parameters of the pre-trained model. This approach generally performs better when a sufficiently large target dataset is available [31].
- **Partial fine-tuning**: Update only a subset of parameters, such as the final layers or certain attention blocks. This is useful when the target dataset is moderate in size, helping to prevent overfitting [81].
- **Feature extraction**: Keep the pre-trained model frozen and use it to extract features from the input. Only a downstream model (e.g., a classifier or a smaller neural network) is trained on these features. This approach works well for small target datasets or when computational resources are limited [82].

2.5. Evaluation Metrics

2.5.1. Word Error Rate

In ASR, the performance of a system is typically evaluated using the Word Error Rate (WER). WER is a standard metric that quantifies the accuracy of transcribed speech by comparing the ASR output with a reference transcription. WER is defined as the ratio of the total number of errors (insertions, deletions, and substitutions) to the total number of words in the reference transcription. Mathematically, WER is calculated as

$$\text{WER} = \frac{S + D + I}{N} * 100\% \tag{2.5}$$

where

- S represents the number of substitutions (incorrect words that are different from reference).
- D represents the number of deletions (words that were in the reference but were missed by the ASR system).
- I represents the number of insertions (extra words added by the ASR system that were not in the reference).
- N is the total number of words in the reference transcription.

2.5.2. Statistical Significance

Matched-Pair Sentence Segment Word Error (MAPSSWE) [83] is a statistical test to assess whether the difference in WER of two ASR systems is statistically significant. In contrast with straightforward WER comparisons, the MAPSSWE procedure conducts a paired comparison of two systems at the sentence level, meaning each utterance is evaluated independently by both systems, and the differences are compared within matched pairs (i.e., the same sentence across systems). This approach ensures that any observed performance differences are not due to random chance but rather to system improvements. It is particularly useful when comparing two ASR models evaluated on the same test set.

The open source tool WER-SigTest [84] was used to determine whether differences in WER in the same test set were statistically significant. This tool includes a script that compares the transcription hypotheses of different ASR systems to perform the significance test. The toolkit calculates the p-value by computing the MAPSSWE [83] between the outputs of different ASR systems. The formulas for calculating the p-value of MAPSSWE are shown below.

The difference in matched pairs for each sentence i is calculated as

$$d_i = e_{i1} - e_{i2} \tag{2.6}$$

where:

- e_{i1} is the number of word errors for system 1 on sentence i,
- e_{i2} is the number of word errors for system 2 on sentence i.

The mean difference between n sentence pairs is:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{2.7}$$

The standard deviation of the differences is:

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2} \tag{2.8}$$

The t-statistic is then computed as:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{2.9}$$

where:

- t is the test statistic,
- n is the number of matched sentence segments.

The resulting t-value is compared against a t-distribution with $n-1$ degrees of freedom to determine the p-value, which indicates the statistical significance of the difference between the two systems. The results' statistical significance is denoted by stars following the WER values: *** indicates a significant difference at the level of p=0.001, ** indicates a significant difference at the level of p=0.01 but not p=0.001, * indicates a significant difference at the level of p=0.05 but not p=0.01, no stars suggests that the performance difference is statistically insignificant, indicating that the observed improvement could have occurred due to random variation.

3 Methodology

In this chapter, a detailed description of the data sets and the training strategies used in this thesis is provided. In Section 3.1, we provide a detailed description of the selected data sets. Two baseline models were introduced in Section 3.2 with their implementation. The implementation of data augmentation methods is described in Section 3.3. The fine-tuning procedures are detailed in Section 3.4. Section 3.5 presents the integration of data augmentation and fine-tuning.

3.1. Datasets

I will introduce all the data sets used in the thesis. Both the Auris and Jasmin datasets are used in data augmentation to generate more simulated child data. Both are also used for fine-tuning models. The CGN data set is only used to train the conformer-based ASR model from scratch, which is our first baseline. Auris and Jasmin datasets are also used as our test sets.

3.1.1. Auris

Auris is a Dutch corpus consisting of speech recordings of conversations between children (3 to 9) with (possible) developmental language disorders and their speech therapists. The given Auris data set originally had 47 audio files and their corresponding configuration files (.TextGrid), including the transcription of each utterance in the audio files and the corresponding timestamps. However, the configuration files of two of the audio files contained incorrect or missing transcriptions and timestamps, so these two audio files were not used in the experiments in this paper. In the end, all experiments in this paper used only 45 audio files, which contained conversations between 40 different target children and their speech therapists. Among these children, five had two audio files each, while the remaining 35 children had one audio file each. There are 5 audio files missing age information, for the purpose of splitting the data by age, these files were treated as belonging to a single “unknown” age group. This setup ensures that all audio files are accounted for while allowing a complete age distribution despite incomplete age metadata.

The Auris data set was processed and segmented to extract only the speech of the target children, excluding the speech of the therapists. To process these files, a Python program using a dedicated library was developed to read and write .TextGrid files. During processing, we encountered errors in a small number of TextGrid files, caused by ”invalid timestamps”, such as intervals ending earlier than they started or overlapping with adjacent intervals. These inconsistencies, likely due to annotation errors, were manually corrected or skipped to ensure a reliable extraction of the child’s speech segments.

Segmentation was necessary because the original audio files are very long, often tens of minutes, which can lead to errors or excessive memory usage when running the ASR models. By extracting only valid child speech segments, the ASR models were trained exclusively on the children’s speech, avoiding irrelevant interference from the therapists’ speech or problematic intervals. This ensures that the models learn speech information only relevant to children, which is crucial for accurately evaluating ASR performance for children with DLD.

To guarantee robust model training and balanced evaluation, the data set was divided into three subsets according to a standard ratio: **70% for training (32 audio recordings), 15% for validation (7 audio recordings), and 15% for testing (6 audio recordings)**. For the 5 children who had 2 audio files each, since the 2 audio recordings were made over a long period of time, both for a few months, we considered the total of 10 audio recordings to be audio recordings of different children.

Furthermore, to maintain demographic consistency and minimize potential bias, we preserved the approximate gender ratio of the entire data set approximately **2.75:1 male to female** within each subset. This ratio reflects the original gender distribution of the data **(33 males and 12 females)** and was carefully replicated across splits to guarantee fair representation. We also consider the age distribution during the data division. There was no intersection of speakers between the training, validation, and test sets. After data splits, it resulted in 4.43 hours of processed training data, 40.59 minutes of validation data, and 48.38 minutes of test data. The overview of the Auris data split is shown in Table 3.1.

The gender composition for each subset is as follows:

- **Training set:** 24 males, 8 females
- **Validation set:** 5 males, 2 females
- **Test set:** 4 males, 2 females

3.1.2. The Spoken Dutch corpus (CGN)

CGN [85], also known as the Corpus Gesproken Nederlands, is a Dutch corpus that consists of native Dutch speech spoken by adults (18–65 years) from the Netherlands and Flanders. We only use the data collected in the Netherlands to train our ASR systems. The corpus includes a wide range of speech types, such as read speech, interviews, debates, and broadcast news. The following Table 3.2 outlines 15 distinct components of speech data, each accompanied by a concise description. Notably,

Table 3.1: Auris Data Splits

Age	Total	Train	Validation	Test
3	1	1M	0	0
4	5	3M	1M	1M
5	18	8M/6F	1M/1F	1M/1F
6	10	7M/1F	1M	1F
7	5	1M/1F	1M/1F	1M
9	1	1M	0	0
Unknown Age (Males)	5	3M	1M	1M
Total	45 (33 male, 12 female)	32	7	6
Durations	-	4.43 hours (A_{train})	40.59 minutes	48.38 minutes (A_{test})
Number of Utterances	-	7058	1254	1153

components labeled a through h relate to multilogue speech data, whereas those identified as i through o correspond to monologue speech data. The unprocessed training data consist of 483 hours. After preprocessing the data, it resulted in 428.58 hours of processed training data C_{train} and 4.35 hours of validation data. The training and validation set partitions adhere to the experimental setup described in [86]. The overview of CGN data split is shown in Table 3.3. There was no intersection of speakers between the training and validation sets. Because in this study, we only want to investigate the ASR performance on children's speech, so the test set of the CGN dataset is not used in the following experiments.

Table 3.2: CGN Components and Descriptions

Component	Description
a	face-to-face spontaneous conversations
b	interviews with teachers of Dutch
c	spontaneous telephone dialogues (switchboard)
d	spontaneous telephone dialogues (local interface)
e	simulated business negotiations
f	interviews/discussions/debates (broadcast)
g	(political) discussions/debates/meetings
h	lessons recorded in a classroom
i	live commentaries (broadcast)
j	newsreports/reportages (broadcast)
k	news (broadcast)
l	commentaries/columns/reviews (broadcast)
m	ceremonious speeches/sermons
n	lectures/seminars
o	read speech

Table 3.3: Total duration and number of utterances for each data split of CGN

Dataset Split	Duration (hours)	Number of Utterances
Training	428.58 (C_{train})	697250
Validation	4.35	7043

3.1.3. Jasmin-CGN

Jasmin-CGN [87] is an extension of the CGN corpus and includes spoken speech by native Dutch speakers of different age groups (children, teenagers, and older adults) as well as non-native speakers of Dutch who are teenagers and adults. Speech includes both read speech (RS) and human-machine interaction (HMI) speech. The general information and the matching duration of the raw speech data for the five speaker groups in the Jasmin-CGN corpus are detailed in Table 3.4 below. Only native Dutch children (DC) aged 6 to 13 years would be used in the following experiments. The target children in the Auris data set are all below 10 years of age. To provide a fair performance comparison between typical and atypical child speech. I will divide the Dutch children's speech into a training set, a validation set, and a test set. The training data set, which comprises 6.63 hours of child speech, called J_{train}, was used to train ASR systems. The validation set comprises 44.22 minutes of child speech from three female and three male Dutch child speakers, with one child of each age from 7 to 12 years. The test set comprises 35.73 minutes of read speech and 8.65 minutes of HMI speech from three female and three male Dutch child speakers, with one child of each age from 7 to 12 years, which is denoted by R_{DC} and H_{DC}. The partition of the test set is based on the setup of the experiment in [19]. The overview of the Jasmin data split is shown in Table 3.5. There was no intersection of speakers between the training, validation, and test sets. The detailed speaker information of the validation set and the test set are shown in Tables 3.6 and 3.7, respectively.

Table 3.4: Overview of Speaker Groups and Data Duration of Jasmin-CGN

Code	Group Description	Age Range	Duration
DC	Native Dutch children	6–13	12h 21m
DT	Native Dutch teenagers	12–18	12h 21m
DOA	Native Dutch older adults	$\geq$ 59	9h 26m
NNT	Non-native teenagers	11–18	12h 21m
NNA	Non-native adults	19–55	12h 21m

Table 3.5: Total duration and number of utterances for each data split of Jasmin-CGN

Dataset Split	Duration	Number of Utterances
Training	6.63 hours (J_{train})	13945
Validation	44.22 minutes	1551
Testing	35.73 minutes (R_{DC}), 8.65 minutes (H_{DC})	1213 (R_{DC}), 350 (H_{DC})

Table 3.6: Speaker information of validation set: Native Dutch children speakers

Speaker ID	Gender	Age
N000026	Male	8
N000028	Male	10
N000030	Female	9
N000050	Female	12
N000062	Female	11
N0000210	Male	7

Table 3.7: Speaker information of test set R_{DC} and H_{DC}: Native Dutch children speakers

Speaker ID	Gender	Age
N000025	Female	8
N000027	Male	9
N000029	Male	10
N000054	Female	11
N000045	Male	12
N0000213	Female	7

3.2. Baselines

Two baseline ASR models were trained to provide a point of comparison, allowing us to evaluate the effects of fine-tuning and data augmentation across different architectures and to ensure that the ob-

served trends were not model-specific. The first baseline model is an encoder-decoder conformer [88] based model implemented using the ESPnet Toolkit [74]. It combines self-attention and convolutional modules to capture both global and local dependencies in the input speech signal. I chose this model because the experimental results in [89] demonstrated that the Conformer-based ASR model outperformed the Transformer-based model used in the earlier study [19] to mitigate bias against non-native Dutch accents when evaluated in the CGN and Jasmin-CGN corpora. Furthermore, there is an increasing trend for using pre-trained models in ASR, and it shows good performance, especially when fine-tuned on some low-resource data set [35, 90]. Therefore, we want to compare this training-from-scratch model with a state-of-the-art (SOTA) pre-trained model. So we chose OpenAI Whisper models [75] as the second baseline since it is a widely used SOTA pre-trained model.
The training data set from the CGN corpus C_{train} was used to build the conformer-based baseline.

3.2.1. Conformer-based ASR model implementation

ASR model configuration

The conformer encoder contains 12 layers, each with 4 attention heads and a position-wise feed-forward layer whose dimensionality is 1024. A dropout of 0.1 was used on the whole encoder to reduce overfitting.

The decoder is based on the Transformer architecture using a hybrid CTC/Attention decoding strategy with CTC weight 0.3 and contains 6 layers, each with 4 attention heads and 1024 feed-forward units. Dropout methods were used the same in the decoder.

The input features were extracted directly from the raw audio using 80-dimensional log-Mel filterbanks. Byte Pair Encoding (BPE) was used as the subword tokenization strategy with a vocabulary size of 5000 units.

The first baseline, the conformer-based ASR model, was trained from scratch for 40 epochs. The final model is chosen on the basis of the average of the top 10 models with the highest accuracy evaluated on the validation set.

3.2.2. Whisper model implementation

For the second baseline, we use the Whisper large-v3 model, as it shows the lowest WER of Dutch speech recognition compared to other Whisper models [79]. This baseline, the OpenAI Whisper large v3 pre-trained model is constructed using openly released code and model weights [91], although the training data itself are not publicly available. '

3.3. Data Augmentation

I used the built-in script provided by ESPnet [74] to implement the speed perturbation function. In our experiments, I generated child-like speech by speed perturbing the typical Dutch child speech data J_{train} (Jasmin) and atypical Dutch child speech A_{train} (Auris) using the perturbation factors {0.9, 1.1}, resulting in two two-fold data augmentations, which doubling each dataset and produced the augmented sets $SP_{J_{\text{train}}}$ and $SP_{A_{\text{train}}}$, respectively. We used the Python library `nlpaug` [92] to generate one-fold VTLP-augmented data separately for the typical and atypical Dutch child speech datasets, resulting in two augmented datasets: one for typical speech and one for atypical speech. Unlike speed perturbation, which applies two fixed factors, VTLP randomly selects a single factor between 0.9 and 1.1 for each utterance, producing only one-fold vtlp-augmented data set.

3.4. Fine-tuning

Two baseline ASR models were used in this study: a Conformer-based model and Whisper-large-v3. The Conformer-based baseline was trained from scratch on the CGN training set (C_{train}) and serves as the ”pre-trained model” for subsequent Conformer-based ASR fine-tuning experiments. Similarly, the Whisper-large-v3 baseline, which is a pre-trained model provided by OpenAI, serves directly as the ”pre-trained model” for the Whisper-large-v3 ASR fine-tuning experiments.

In this study, fine-tuning allows the ASR models to better recognize child speech while leveraging the general acoustic and linguistic knowledge captured during pre-training.

For the Conformer-based ASR fine-tuning experiments, the corresponding pre-trained Conformer model was continued to be trained for 40 epochs using the same architecture as the baseline model.

This approach updates the model parameters gradually to improve performance on the child speech data while maintaining the general representations learned from the CGN adult speech.

For the Whisper-large-v3 ASR fine-tuning experiments, the pre-trained Whisper model was fine-tuned for 4000 training steps. During this process, the WER was evaluated every 1000 steps in the validation sets and the final model was selected based on the best WER achieved in the validation data.

3.5. Combined Data Augmentation and Fine-tuning

In addition to performing data augmentation and fine-tuning separately, this study also investigates the effect of combining augmented child speech data with fine-tuning on ASR performance. By integrating augmented datasets with the fine-tuning process, the ASR models are exposed to a more diverse set of child speech variations while retaining the general acoustic and linguistic knowledge acquired during pre-training.

For the Conformer-based ASR experiments, the corresponding pre-trained Conformer model was fine-tuned for 40 epochs using the same architecture as the baseline. The fine-tuning data consisted of the Jasmin-CGN child speech training set (J_{train}) combined with its augmented versions, or the Auris training set (A_{train}) combined with its augmented versions. This approach allows the model parameters to adapt to the effect introduced by the augmented child speech while maintaining the general representations learned from the original child speech.

For the Whisper-large-v3 ASR experiments, the pre-trained Whisper model was fine-tuned for 4000 steps, following the same procedure as described for the Conformer-based experiments mentioned in the previous paragraph. During this process, the WER was evaluated in the validation sets every 1000 steps and the final model was selected based on the lowest WER on the validation data.

4
Experiments

This section describes the experimental design used to investigate the three research questions outlined in Section 1.2. The experiments are grouped by research question, with references to the methods described in Section 3.

4.1. Baselines

This first baseline model (conformer-based model) is trained from scratch using

- CGN training set C_{train}

The second baseline model (Whisper-large-v3) is directly evaluated without any additional training. All models explained in this and the following subsections are evaluated on the same test datasets:

- RS and HMI test sets of the Jasmin-CGN child corpus (R_{DC} and H_{DC}); Auris test set (A_{test}) using WER as a metric

4.2. RQ1: Data Augmentation

To answer RQ1, a series of experiments were conducted. The augmentation techniques applied included SP and VTLP, both individually and in combination. SP modifies the speaking rate by simulating natural variations in tempo among children, while VTLP modifies the spectral characteristics by simulating changes in vocal tract length. Applying them individually allows assessing their separate effects on ASR performance, whereas combining them means merging the SP-augmented and VTLP-augmented datasets, without applying both augmentations sequentially to the same data. This tests whether combining temporal and spectral variations provides an additional effect. Since Whisper is a pre-trained model, we can only examine its fine-tuning performance. Therefore, experiments will only be conducted using the first baseline model. Specifically, the conformer-based baseline model was trained for 40 epochs using the same architecture as the baselines. All experiments include original CGN training data (C_{train}) combined with original child speech training data from the Jasmin-CGN corpus (J_{train}) or from the Auris corpus (A_{train}). In some configurations, the original child speech training data are further combined with their augmented versions. The configurations differ in the type of child data and augmentation applied:

- The SP-augmented training set of Jasmin-CGN child speech $SP_{J_{\text{train}}}$
- The VTLP-augmented training set of Jasmin-CGN child speech $VTLP_{J_{\text{train}}}$
- The SP- and VTLP-augmented training set of Jasmin-CGN child speech $SP_{J_{\text{train}}} + VTLP_{J_{\text{train}}}$
- The SP-augmented training set of Auris child speech $SP_{A_{\text{train}}}$
- The VTLP-augmented training set of Auris child speech $VTLP_{A_{\text{train}}}$
- The SP- and VTLP-augmented training set of Auris child speech $SP_{A_{\text{train}}} + VTLP_{A_{\text{train}}}$

4.3. RQ2: Fine-Tuning

To answer RQ2, both baseline models, conformer-based and Whisper-large-v3, were fine-tuned on the original training set of the Jasmin-CGN child corpus J_{train} or the original training set of the Auris corpus A_{train}.

- The training set of Jasmin CGN child speech J_{train}
- The Auris training set A_{train}

4.4. RQ3: Integration of Data Augmentation and Fine-Tuning

To answer RQ3, all experiments were conducted using the same training strategy as used in Section 3.4. Both baseline models were fine-tuned in the original training set of the Jasmin-CGN child corpus (J_{train}) or that of the Auris corpus (A_{train}), each combined with its respective augmented data. The differences between models lie in the type of augmented data included:

- The SP-augmented training set of Jasmin-CGN child corpus $SP_{J_{\text{train}}}$
- The VTLP-augmented training set of Jasmin-CGN child corpus $VTLP_{J_{\text{train}}}$
- The SP- and VTLP-augmented training set of Jasmin-CGN child corpus $SP_{J_{\text{train}}}$ + $VTLP_{J_{\text{train}}}$
- The SP-augmented Auris training set $SP_{A_{\text{train}}}$
- The VTLP-augmented Auris training set $VTLP_{A_{\text{train}}}$
- The SP- and VTLP-augmented Auris training set $SP_{A_{\text{train}}}$ + $VTLP_{A_{\text{train}}}$

5
Results

This chapter begins with a description of the experimental results for the two baselines in Section 5.1. RQ1 is explored through data augmentation experiments, the results of which are detailed in Section 5.2. For RQ2, fine-tuning experiments are conducted and discussed in Section 5.3. Section 5.4 addresses RQ3 by integrating data augmentation with fine-tuning and presenting the corresponding findings.

5.1. Baselines

Before addressing the main RQ, it is essential to first establish the baseline performance in both typical and atypical child speech. The baseline speech recognition performance, measured by WER, is summarized in Table 5.1 under the rows labeled **BL1: Conformer-based** and **BL2: Whisper-large-v3** in the **Model** column.

First, for the row with ”BL1: Conformer-based” in Table 5.1, the model was trained from scratch on C_{train}. In the Auris test set, the conformer-based baseline shows a considerably high WER, while the performance in the Jasmin-CGN test sets is slightly better for read speech (R_{DC}) than for HMI speech (H_{DC}). This shows that the model works better on read speech than on HMI-type speech, which also corresponds to expectations, since read speech is usually more acoustically clean and controlled.

Second, for the row with ”BL2: Whisper-large-v3”, which is a pre-trained model without using C_{train}, the WER on the Auris test set is 70.9%, showing a large improvement over the conformer-based baseline despite having no additional training. For the Jasmin-CGN test sets, Whisper also achieves lower WERs in both subsets, performance is better in read speech (R_{DC}) than in HMI speech (H_{DC}), again confirming a better performance in read speech than in HMI speech. As shown in Table 5.1, the Whisper-large-v3 model achieves markedly better performance than the Conformer-based model in all test cases.

5.2. Data Augmentation

I merged C_{train} and J_{train} to retrain the conformer-based model, and then retrain the conformer-based model with C_{train} and J_{train} and augmented data from J_{train}. The results are shown in rows where the **“Model”** column is labeled **Conformer-based** in Table 5.1. The WER results for all test sets are shown with significance test values in brackets following each WER. If only one significance result is reported, it represents a comparison with the baseline model. If an additional significance result appears in brackets following the first, it indicates a comparison with the model trained on the original child data. The stars (*, **, ***) denote statistically significant improvements, whereas the empty brackets () indicate that there is no significant difference.

All these results for both the Auris and Jasmin test sets show significant improvement compared to the baseline model (BL1). However, these comparisons do not distinguish whether the improvement arises from simply adding more child speech data or from the inclusion of augmented data. Therefore, the focus is placed on comparisons against the model trained with only the original child data to isolate the pure effect of augmentation.

Using Jasmin Child Speech Training Data

Compared to the model trained with C_{train} + J_{train}, both SP and SP + VTLP produced statistically significant reductions in WER in the Auris test set ($p < 0.001$) and in the Jasmin read-speech test set ($p < 0.001$). However, the improvements in HMI speech were not statistically significant, indicating that there is no effect on performance. VTLP also significantly improved performance compared to C_{train} + J_{train}, but to a lesser extent in the Auris test set ($p < 0.05$), and significantly improved performance in HMI speech ($p < 0.05$).

Using Auris Child Speech Training Data

Compared to the model trained with $C_{train}+A_{train}$, only SP produced a statistically significant reduction in WER in the Auris test set ($p < 0.05$). The improvements observed in the Jasmin test sets were not statistically significant, indicating that there is no reliable effect of SP augmentation on typical child speech. VTLP did not produce statistically significant changes in all test sets, suggesting that VTLP alone had little impact. In contrast, the combination of SP + VTLP did not significantly improve Auris performance and even resulted in significant performance reductions in both the Jasmin read-speech test ($p < 0.05$) and the HMI test ($p < 0.001$).

5.3. Fine-tuning

Using Jasmin Child Speech Training Data

Both baselines were fine-tuned using Jasmin child training data J_{train}, and their corresponding results are presented in Table 5.1 under the title: **Using Jasmin Child Speech Training Data**. Specifically, these results appear in the two rows where the **"Model"** column is labeled **Fine-tuning** and the **"Train data"** column indicates J_{train}.

When fine-tuned on the Jasmin dataset (J_{train}), the Conformer-based model showed a statistically significant improvement over its baseline performance on both datasets ($p < 0.001$). In the Auris test set (A_{test}), WER decreased by approximately 7%, while in the Jasmin test sets, the reductions were much more noticeable around 30% for read speech (R_{DC}) and 25% for the HMI condition (H_{DC}), demonstrating a strong positive effect of fine-tuning. The Whisper-based model demonstrated a similar trend, with statistically significant improvements of approximately 20% in both Jasmin test sets ($p < 0.001$). In contrast, the small increase in WER observed in the Auris test set was not statistically significant, indicating that fine-tuning J_{train} did not reliably improve performance in the Auris data.

Using Auris Child Speech Training Data

Fine-tuning with the Auris dataset also resulted in a statistically significant reduction in WER for both baseline models ($p < 0.001$). For BL1: Conformer-based model, fine-tuning produced an approximate large relative reduction of 30% in WER in the Auris test set, while WER in the Jasmin test sets decreased much less, approximately 15%. Similarly, the Whisper-based model (BL2) showed similar significant improvements with reductions of approximately 15% on the Auris test set and a much smaller reduction of approximately 5% on the Jasmin test sets ($p < 0.001$).

5.4. Combined Data Augmentation and Fine-tuning

Using Jasmin Child Speech Training Data

Both baselines are fine-tuned using original Jasmin child training data and its augmented data J_{train}, and their corresponding results are presented in Table 5.1 under the title: **Using Jasmin Child Speech Training Data**. Specifically, these results appear in the rows under where the **"Model"** column is labeled **Fine-tuning** and the **"Train data"** column indicates J_{train} **+ its augmented data**. These results are compared directly with the baseline models (BL1 and BL2) to evaluate the overall combined effect of fine-tuning and data augmentation.

When fine-tuned with the original Jasmin training set and its augmented data, the conformer-based model (BL1) achieved statistically significant improvements in both the Auris and Jasmin test set ($p < 0.001$) compared to its baseline. Among the three combinations of augmentation, VTLP achieved the lowest WER in Auris, while SP and SP+VTLP also provided consistent improvements. However, VTLP produced the highest WER in read speech and a slightly higher WER in HMI speech. In contrast, SP + VTLP achieved the lowest WER in read speech but the highest WER in HMI speech. SP alone provided the lowest WER on HMI speech while maintaining competitive performance on read speech. For the Whisper-based model (BL2), fine-tuning with the original Jasmin training set and its augmented data produced a higher WER in the Auris test set A_{test}, which exhibited a significant decrease in performance ($p < 0.05$). However, it achieved a significant improvement, producing the lowest average WERs in both Jasmin test sets ($p < 0.001$).

Using Auris Child Speech Training Data

When fine-tuned with the original Auris training set and its augmented data, both models achieved significantly lower WER on the Auris test set A_{test} ($p < 0.001$) compared to their baselines. The conformer-based model (BL1) showed a larger improvement of approximately 30%, while the Whisper-based model (BL2) achieved a much smaller improvement of about 15%. In the Jasmin test sets, BL1 (conformer-based model) also exhibited statistically significant improvements of roughly 15% in both read and HMI speech. For BL2 (Whisper-based model), only the model fine-tuned with VTLP demonstrated significant improvements in both the read speech test set R_{DC} ($p < 0.001$) and the HMI test set H_{DC} ($p < 0.01$) while the models fine-tuned with SP and SP + VTLP showed no statistically significant changes on the HMI test, indicating that they did not reliably improve performance in the HMI test set H_{DC}.

Table 5.1: WER results and significance test results on Auris and Jasmin datasets. All the numbers with bold means the best result in each column

Details		**Auris (% WER)**	**Jasmin (% WER)**	
Model	**Train data**	A_{test}	R_{DC}	H_{DC}
BL1: Conformer-based	C_{train}	98.1	43.1	45.4
BL2: Whisper-large-v3	–	70.9	25.6	35.1
Using Jasmin Child Speech Training Data				
Conformer-based	C_{train} + J_{train}	90.3***	14.2***	19.5***
Conformer-based	C_{train} + J_{train} + $SP_{J_{\text{train}}}$	87.1***(***)	11.7***(***)	18.7***()
Conformer-based	C_{train} + J_{train} + $VTLP_{J_{\text{train}}}$	88.3***(*)	12.3***(***)	17.4***(*)
Conformer-based	C_{train} + J_{train} + $SP_{J_{\text{train}}}$ + $VTLP_{J_{\text{train}}}$	87.4(***)	9.7***(***)	18.3***()
Fine-tuning: Conformer-based	J_{train}	90.7***	10.6***	18.8***
Fine-tuning: Conformer-based	J_{train} + $SP_{J_{\text{train}}}$	89.2***	8.5***	17.7***
Fine-tuning: Conformer-based	J_{train} + $VTLP_{J_{\text{train}}}$	88.1***	9.2***	18.7***
Fine-tuning: Conformer-based	J_{train} + $SP_{J_{\text{train}}}$ + $VTLP_{J_{\text{train}}}$	88.8***	7.6***	19.6***
Fine-tuning: Whisper-large-v3	J_{train}	79.1	6.4***	15.1***
Fine-tuning: Whisper-large-v3	J_{train} + $SP_{J_{\text{train}}}$	85.9*	**5.7*** **	**14.3*** **
Fine-tuning: Whisper-large-v3	J_{train} + $VTLP_{J_{\text{train}}}$	84.6*	5.9***	14.4***
Fine-tuning: Whisper-large-v3	J_{train} + $SP_{J_{\text{train}}}$ + $VTLP_{J_{\text{train}}}$	94.8*	**5.7*** **	15.3***
Using Auris Child Speech Training Data				
Conformer-based	C_{train} + A_{train}	66.6***	25.0***	27.5***
Conformer-based	C_{train} + A_{train} + $SP_{A_{\text{train}}}$	64.7***(*)	24.8***()	28.2***()
Conformer-based	C_{train} + A_{train} + $VTLP_{A_{\text{train}}}$	65.8***()	24.6***()	28.6***()
Conformer-based	C_{train} + A_{train} + $SP_{A_{\text{train}}}$ + $VTLP_{A_{\text{train}}}$	65.1***()	26.4***(*)	32.2***(***)
Fine-tuning: Conformer-based	A_{train}	67.7***	26.8***	32.3***
Fine-tuning: Conformer-based	A_{train} + $SP_{A_{\text{train}}}$	64.6***	27.7***	31.7***
Fine-tuning: Conformer-based	A_{train} + $VTLP_{A_{\text{train}}}$	65.9***	27.0***	30.2***
Fine-tuning: Conformer-based	A_{train} + $SP_{A_{\text{train}}}$ + $VTLP_{A_{\text{train}}}$	64.6***	27.9***	32.3***
Fine-tuning: Whisper-large-v3	A_{train}	54.1***	20.5***	29.5***
Fine-tuning: Whisper-large-v3	A_{train} + $SP_{A_{\text{train}}}$	53.6***	21.7***	32.6
Fine-tuning: Whisper-large-v3	A_{train} + $VTLP_{A_{\text{train}}}$	**53.2*** **	21.2***	30.6**
Fine-tuning: Whisper-large-v3	A_{train} + $SP_{A_{\text{train}}}$ + $VTLP_{A_{\text{train}}}$	54.7***	23.4**	32.9

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

6

Discussions and Conclusions

In this chapter, the results presented in Section 5 are analyzed and discussed in detail. Section 6.1 discusses the answers to individual research questions: the findings of RQ1, RQ2, and RQ3 are described in Section 6.1.1, Section 6.1.2, and Section 6.1.3. Finally, in Section 6.1.4, I present a more general reflection relating to the main research question. Section 6.2 gives an overview of the study. Finally, in Section 6.3, I present possible future directions of this research based on results and limitations in this work.

6.1. Discussion

6.1.1. Implications of Data Augmentation

This section discusses the effect of data augmentation techniques, SP and VTLP, on ASR performance. For the conformer-based model trained using Jasmin child data, both SP and SP + VTLP showed clear improvements in the Auris and Jasmin read speech. However, improvements in HMI speech were limited, indicating that improvement is less effective for spontaneous or conversational speech. VTLP showed moderate improvements compared to the model trained with (C_{train} + J_{train}), showing smaller improvements in the Auris test set but considerable benefits for HMI speech.

When the conformer-based model was trained using Auris child data, the benefits of augmentation were more constrained. Only SP produced a clear improvement on the Auris test set, with a limited effect on the Jasmin test set, while VTLP alone had a small impact on both the Auris and Jasmin test sets. Moreover, combining SP and VTLP led to a small improvement on the Auris test set, although the effect was not meaningful, and even resulted in considerably reduced recognition performance on the Jasmin test sets, particularly for HMI speech.

In general, these findings demonstrate that data augmentation methods clearly improve the performance of ASR for child speech with DLD (Auris) but in some cases reduce the accuracy of typical child speech (Jasmin), which is in line with other research focused on dysarthric speech [27, 29], particularly when applying SP, which consistently provided the largest reductions in WER. VTLP produced smaller, but overall consistent effects. The combined SP + VTLP method was most effective when trained on Jasmin data, but less reliable and even led to a noticeable reduction in the recognition accuracy of the Jasmin test set when trained on Auris data.

6.1.2. Implications of Fine-tuning

This section discusses the effect of transfer learning through fine-tuning on ASR performance. The results presented in Section 5.3 show that fine-tuning substantially improves recognition of speech from children with DLD (Auris) while maintaining or improving performance on typical child speech (Jasmin).

When fine-tuned on typical child speech (Jasmin), the conformer-based model (BL1) achieved considerable improvements in performance in both atypical (Auris) and typical child speech (Jasmin). In the case of Whisper, fine-tuning using original Jasmin training data appeared to reduce its ability to recognize atypical child speech (Auris) but still improve performance substantially on Jasmin data. However, this observed performance reduction was minor and not considered meaningful. This indicates a possibility that Whisper's internal representations, which are learned from a wide variety of data, cannot recognize atypical child speech well when fine-tuning Jasmin data and thus could result in reduced performance for atypical child speech.

Fine-tuning on the Auris dataset (atypical child speech) produced the strongest improvements in atypical child speech (Auris) recognition for both models. Whisper achieved the lowest WER, confirming its strong adaptation capabilities even when fine-tuned on a relatively small dataset, while the conformer-based model also showed clear improvements under the same conditions.

In particular, fine-tuning the models using Auris data even improves the performance significantly on the Jasmin test sets for both models, suggesting that there was no large negative transfer across domains when fine-tuning. Although the absolute reductions in WER in Jasmin were smaller than those in Auris, the results indicate that fine-tuning with atypical child speech contributes to improved recognition in both types of speech.

In conclusion, these results indicate the effectiveness of fine-tuning to adapt ASR models to child speech. When applied with matched training data, such as fine-tuning atypical child speech for atypical child speech (DLD) recognition, fine-tuning leads to substantial performance improvements. When applied with unmatched data, such as fine-tuning on typical child speech, the models still show noticeable improvements, though to a lesser extent. In general, fine-tuning ASR models using child speech improves the performance substantially to recognize DLD child speech while also improving the performance on typical child speech, which is in line with other research focused on child speech [35].

6.1.3. Implications of combining Data Augmentation and Fine-tuning

This section discusses the combined effect of data augmentation and fine-tuning on ASR performance by comparing the fine-tuned models trained with augmented data directly against the baseline systems (BL1 and BL2). This approach provides a clear measure of the overall improvement achieved when both techniques are applied together. In particular, we examine whether combining these methods improves recognition of speech from children with DLD while maintaining accuracy in typical child speech.

When Jasmin child speech (and its augmentations) is used for fine-tuning, the conformer-based model showed noticeable improvements in both DLD and typical child speech, but the effect depends on speech style. VTLP is most effective for Auris data, whereas SP resulted in the best outcomes for HMI speech, combining SP + VTLP produced the lowest read speech WER but the highest WER of HMI. In contrast, Whisper improved the recognition performance considerably on Jasmin test data and reached the average best WERs, but it degraded noticeably on Auris, suggesting that augmenting typical child speech does not transfer well to DLD child speech for a heavily pre-trained model.

Using Auris (and its augmentations) as fine-tuning data produced substantial recognition performance for both models, with a larger effect on the conformer-based model than on the Whisper-based model. In typical speech (Jasmin), the conformer showed clear recognition performance in read and HMI indicates that fine-tuning using DLD child speech can be transferred positively to typical child speech. However, for Whisper, only VTLP transfers well to Jasmin (read and HMI). SP and SP+VTLP have little benefit in HMI, suggesting that temporal perturbations are less compatible with Whisper's pre-trained representations for spontaneous speech, but its average recognition performance in Auris is the best.

Combining augmentation with fine-tuning improves DLD recognition substantially when the fine-tuning data are matched (Auris to Auris), and moderately to not at all when mismatched (Jasmin to Auris). Typical child speech accuracy is improved for BL1, for BL2 it is maintained or improved.

6.1.4. General Discussion

This section discusses how ASR performance can be improved for the speech of children with DLD without negatively affecting the recognition accuracy in typical child speech.

For the conformer-based ASR model, the results show that the best way to improve DLD child speech recognition performance is to combine data augmentation with fine-tuning when domain-matched DLD child speech (Auris) is used. Both SP and SP + VTLP augmentations showed the best WERs in child speech with DLD. Importantly, typical child speech (Jasmin) performance is improved, confirming that fine-tuning using DLD child speech does not lead to degradation in typical child speech recognition.

For the Whisper-based ASR model, the combined approach also improves DLD speech recognition, although the improvements are more moderate due to the model's strong pre-trained representations. Fine-tuning on domain-matched data (Auris to Auris) leads to the best performance, while mismatched fine-tuning (Jasmin to Auris) showed a noticeably recognition performance reduction. Typical child speech accuracy remains stable or substantially improved. Whisper fine-tuned on SP-augmented Jasmin data and original Jasmin data, produced its best performance on Jasmin data. This best-performing configuration outperforms previously reported results on similar Jasmin speech recognition tasks using the same Jasmin test set. Previous works, for example, indicate WERs of 6.4% on read speech and 16.6% on HMI speech for the Jasmin test set [89]. However, this improvement comes at the cost of the reduced recognition accuracy of DLD child speech.

In general, the results demonstrate that the combination of data augmentation and fine-tuning using the large pre-trained model provides an effective way to improve ASR performance for children with DLD while maintaining accuracy in typical child speech. The approach enhances model generalization across diverse child speech domains and confirms that domain-matched fine-tuning with augmented data yields the most reliable improvements.

6.2. Conclusion

The aim of this thesis is to investigate how ASR performance can be improved for the speech of children with DLD, while also preserving high accuracy rates in typical child speech. To address this, two SOTA ASR models (a conformer-based model and Whisper-large-v3) were evaluated under multiple training conditions, including baseline models, transfer learning by fine-tuning and data augmentation techniques, such as SP and VTLP.

The results clearly demonstrated that fine-tuning using speech from children with DLD substantially improves the recognition performance when testing on the same domain data (DLD). The use of SP and VTLP in data augmentation also contributed to a lower WER. The best performance among all experiments (53.2% WER in the Auris test set) was achieved by fine-tuning the Whisper model on the Auris training and its VTLP-augmented data. The best performance among all experiments (5.7% WER in read speech and 14.3% WER in HMI speech of Jasmin) was obtained by fine-tuning the Whisper model in Jasmin training and its SP-augmented data. Importantly, this enhanced recognition of atypical child speech did not appear to be at the expense of recognizing typical child speech. Both models retained high levels of performance on typical child speech from the Jasmin dataset, with WERs better than their baselines. However, when fine-tuning was conducted with mismatched data (Jasmin to Auris), the improvement in Auris WER was considerably smaller. This effect was particularly pronounced for the Whisper model, which even showed a notable performance reduction under such mismatched conditions.

In conclusion, this thesis shows that it is possible to improve performance by recognizing speech from children with DLD while maintaining the performance of speech from typical children by using a combination of fine-tuning and data augmentations.

6.3. Future Work

Although this thesis has shown that, combined with domain-specific fine-tuning of DLD speech ASR with data augmentations, the performance for DLD speech is considerably improved, there are definitely some roads to explore forward. There is also a lot of promise with user-adaptive ASR models. Using speaker-adaptive training [93], future work could yield models that are personalized to individual children, which may prove to be especially useful for children with severe or uncommon speech disorders. More advanced augmentation strategies beyond SP and VTLP are also available for testing in future work. Applying techniques such as SpecAugment [94], pitch shifting [95], or using generative adversarial networks (GANs) can enhance robustness and generalizability [96]. SpecAugment applies time and frequency masking to spectrograms, pitch shifting simulates differences in vocal characteristics between children, and GANs generate synthetic child speech. In addition, other transfer learning methods can be applied, such as multi-task learning (MTL) [97], domain adversarial training (DAT) [69]. MTL allows the model to learn multiple related tasks simultaneously, capturing features that benefit recognition of both typical and atypical speech, while DAT encourages the model to learn features that are insensitive to differences between typical and atypical child speech.

References

[1] Larwan Berke, Christopher Caulfield, and Matt Huenerfauth. “Deaf and hard-of-hearing perspectives on imperfect automatic speech recognition for captioning one-on-one meetings”. In: *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 2017, pp. 155–164.

[2] Thomas Kasakowskij and Joerg M Haake. “T3 Talk2Text–A model for near real-time voice transcription in virtual group meetings”. In: *Discover Education* 4.1 (2025), p. 146.

[3] Mike Wald. “Using automatic speech recognition to enhance education for all students: Turning a vision into reality”. In: *Proceedings Frontiers in Education 35th Annual Conference*. IEEE. 2005, S3G–S3G.

[4] Raphael Tang, Ferhan Ture, and Jimmy Lin. “Yelling at your TV: An analysis of speech recognition errors and subsequent user behavior on entertainment systems”. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2019, pp. 853–856.

[5] Vivek Bhardwaj et al. “Automatic speech recognition (asr) systems for children: A systematic literature review”. In: *Applied Sciences* 12.9 (2022), p. 4419.

[6] S Shahnawazuddin et al. “Developing children’s ASR system under low-resource conditions using end-to-end architecture”. In: *Digital Signal Processing* 146 (2024), p. 104385.

[7] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. “Acoustics of children’s speech: Developmental changes of temporal and spectral parameters”. In: *The Journal of the Acoustical Society of America* 105.3 (1999), pp. 1455–1468.

[8] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. “Analysis of children’s speech: Duration, pitch and formants”. In: *Fifth European Conference on Speech Communication and Technology*. 1997.

[9] Abhijit Mohanta and Vinay Kumar Mittal. “Analysis and classification of speech sounds of children with autism spectrum disorder using acoustic features”. In: *Computer Speech & Language* 72 (2022), p. 101287.

[10] Zengjie Yang et al. “Long-term average spectra analysis of voice in children with cleft palate”. In: *Journal of Voice* 32.3 (2018), pp. 285–290.

[11] Benjamin G Schultz et al. “Automatic speech recognition in neurodegenerative disease”. In: *International Journal of Speech Technology* 24.3 (2021), pp. 771–779.

[12] Katrin Tomanek et al. “On-device personalization of automatic speech recognition models for disordered speech”. In: *arXiv preprint arXiv:2106.10259* (2021).

[13] TA Mariya Celin, P Vijayalakshmi, and T Nagarajan. “Data augmentation techniques for transfer learning-based continuous dysarthric speech recognition”. In: *Circuits, Systems, and Signal Processing* 42.1 (2023), pp. 601–622.

[14] Heejin Kim et al. “Dysarthric speech database for universal access research.” In: *Interspeech*. Vol. 2008. 2008, pp. 1741–1744.

[15] Zhengjun Yue. “Continuous speech recognition for people with dysarthria”. PhD thesis. University of Sheffield, 2022.

[16] Craig Fleming et al. “Screening for abdominal aortic aneurysm: a best-evidence systematic review for the US Preventive Services Task Force”. In: *Annals of internal medicine* 142.3 (2005), pp. 203–211.

[17] Dorothy VM Bishop et al. “Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology”. In: *Journal of child psychology and psychiatry* 58.10 (2017), pp. 1068–1080.

[18] Dorothy VM Bishop et al. "CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development. Phase 2". In: *Journal of Child Psychology and Psychiatry* (2017).

[19] Yuanyuan Zhang et al. "Mitigating bias against non-native accents." In: *Interspeech*. 2022, pp. 3168–3172.

[20] Dalia Ritvo et al. "Privacy and Children's Data-An Overview of the Children's Online Privacy Protection Act and the Family Educational Rights and Privacy Act". In: *Berkman Center Research Publication* 23 (2013).

[21] Paul Voigt and Axel Von dem Bussche. "The eu general data protection regulation (gdpr)". In: *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10.3152676 (2017), pp. 10–5555.

[22] David A Van Dyk and Xiao-Li Meng. "The art of data augmentation". In: *Journal of Computational and Graphical Statistics* 10.1 (2001), pp. 1–50.

[23] Siyuan Feng et al. "Quantifying bias in automatic speech recognition". In: *arXiv preprint arXiv:2103.15122* (2021).

[24] Takashi Fukuda et al. "Data Augmentation Improves Recognition of Foreign Accented Speech." In: *Interspeech*. September. 2018, pp. 2409–2413.

[25] Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury. "Data augmentation for deep neural network acoustic modeling". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.9 (2015), pp. 1469–1477.

[26] Tom Ko et al. "Audio augmentation for speech recognition." In: *Interspeech*. Vol. 2015. 2015, p. 3586.

[27] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu. "Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition." In: *Interspeech*. 2018, pp. 471–475.

[28] Tanvina Patel and Odette Scharenborg. "Improving End-to-End Models for Children's Speech Recognition". In: *Applied Sciences* 14.6 (2024), p. 2353.

[29] Mengzhe Geng et al. "Investigation of data augmentation techniques for disordered speech recognition". In: *arXiv preprint arXiv:2201.05562* (2022).

[30] Vishwanath Pratap Singh et al. "Spectral Modification Based Data Augmentation For Improving End-to-End ASR For Children's Speech". In: *arXiv preprint arXiv:2203.06600* (2022).

[31] Hugging Face. *Fine-tuning the ASR model*. `https://huggingface.co/learn/audio-course/en/chapter5/fine-tuning`. 2024.

[32] Daniel V Smith et al. "Improving Child Speech Disorder Assessment by Incorporating Out-of-Domain Adult Speech." In: *Interspeech*. 2017, pp. 2690–2694.

[33] Emre Yilmaz et al. "Combining non-pathological data of different language varieties to improve DNN-HMM performance on pathological speech". In: (2016).

[34] Basil Abraham, Srinivasan Umesh, and Neethu Mariam Joy. "Overcoming Data Sparsity in Acoustic Modeling of Low-Resource Language by Borrowing Data and Model Parameters from High-Resource Languages." In: *INTERSPEECH*. 2016, pp. 3037–3041.

[35] Rishabh Jain et al. "Adaptation of Whisper models to child speech recognition". In: *arXiv preprint arXiv:2307.13008* (2023).

[36] Yixuan Zhang et al. "Comparing data augmentation and training techniques to reduce bias against non-native accents in hybrid speech recognition systems". In: *Proc. 1st workshop on speech for social good (S4SG)*. 2022, pp. 15–19.

[37] Heidi Christensen et al. "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech". In: *Proc. Interspeech 2013*. ISCA, 2013.

[38] Nancy E Hall. "Developmental language disorders". In: *Seminars in Pediatric Neurology*. Vol. 4. WB SAUNDERS COMPANY. 1997, pp. 77–85.

[39] Dorothy V. M. Bishop et al. “CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children”. In: *Journal of Child Psychology and Psychiatry* 58.10 (2017), pp. 1068–1080. DOI: `10.1111/jcpp.12721`.

[40] Marjolijn Van Weerdenburg, Ludo Verhoeven, and Hans Van Balkom. “Towards a typology of specific language impairment”. In: *Journal of Child Psychology and Psychiatry* 47.2 (2006), pp. 176–189.

[41] Gina Conti-Ramsden and Nicola Botting. “Classification of children with specific language impairment: Longitudinal considerations”. In: *Journal of Speech, Language, and Hearing Research* 42.5 (1999), pp. 1195–1204.

[42] Dorothy VM Bishop et al. “CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children”. In: *PLOS one* 11.7 (2016), e0158753.

[43] Susan Rvachew and Françoise Brosseau-Lapré. *Developmental phonological disorders: Foundations of clinical practice*. Plural Publishing, 2016.

[44] Edward S Klein and Cari B Flint. “Measurement of intelligibility in disordered speech”. In: *Measurement* (2006).

[45] Alan G Kamhi et al. “Phonological and spatial processing abilities in language-and reading-impaired children”. In: *Journal of Speech and Hearing Disorders* 53.3 (1988), pp. 316–327.

[46] Laurence B. Leonard. *Children with Specific Language Impairment*. 2nd ed. MIT Press, 2014.

[47] Dorothy VM Bishop. “Comprehension of spoken, written and signed sentences in childhood language disorders”. In: *Journal of Child Psychology and Psychiatry* 23.1 (1982), pp. 1–20.

[48] Laurence B Leonard and Patricia Deevy. “Tense and aspect in sentence interpretation by children with specific language impairment”. In: *Journal of Child Language* 37.2 (2010), pp. 395–418.

[49] Mabel L Rice and John V Bode. “GAPS in the verb lexicons of children with specific language impairment”. In: *First language* 13.37 (1993), pp. 113–131.

[50] Alyssa Kuiack and Lisa Archibald. “Developmental Language Disorder: The childhood condition we need to start talking about”. In: *Frontiers for Young Minds* 7 (2019), p. 94.

[51] Diane J German. “Word-finding intervention for children and adolescents”. In: *Topics in language disorders* 13.1 (1992), pp. 33–50.

[52] Shelley Gray. “Word learning by preschoolers with specific language impairment”. In: *Journal of Speech, Language, and Hearing Research* 47.5 (2004), pp. 1117–1132.

[53] Catherine Adams. “Clinical diagnostic and intervention studies of children with semantic’pragmatic language disorder”. In: *International Journal of Language & Communication Disorders* 36.3 (2001), pp. 289–305.

[54] Heather KJ Van der Lely. “Narrative discourse in Grammatical specific language impaired children: a modular language deficit?” In: *Journal of child language* 24.1 (1997), pp. 221–256.

[55] Susan E Gathercole. “Word learning in language-impaired children”. In: *Child language teaching and therapy* 9.3 (1993), pp. 187–199.

[56] James W Montgomery. “Information processing and language comprehension in children with specific language impairment”. In: *Topics in Language Disorders* 22.3 (2002), pp. 62–84.

[57] Alyssa Kuiack and Lisa Archibald. “Developmental Language Disorder: The childhood condition we need to start talking about”. In: *Frontiers for Young Minds* 7 (2019), p. 94.

[58] Lawrence R Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[59] Geoffrey Hinton et al. “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups”. In: *IEEE Signal processing magazine* 29.6 (2012), pp. 82–97.

[60] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. “Speech recognition with deep recurrent neural networks”. In: *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee. 2013, pp. 6645–6649.

[61] Awni Hannun et al. “Deep speech: Scaling up end-to-end speech recognition”. In: *arXiv preprint arXiv:1412.5567* (2014).

[62] Lawrence R Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. In: *Proceedings of the IEEE* 77.2 (2002), pp. 257–286.

[63] Dong Yu and Lin Deng. *Automatic speech recognition*. Vol. 1. Springer, 2016.

[64] Dario Amodei et al. “Deep speech 2: End-to-end speech recognition in english and mandarin”. In: *International conference on machine learning*. PMLR. 2016, pp. 173–182.

[65] Dilin Wang et al. “Noisy training improves e2e asr for the edge”. In: *arXiv preprint arXiv:2107.04677* (2021).

[66] Bo Li et al. “Towards fast and accurate streaming end-to-end ASR”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 6069–6073.

[67] Bo Li et al. “Scaling end-to-end models for large-scale multilingual asr”. In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 2021, pp. 1011–1018.

[68] Alex Graves et al. “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks”. In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 369–376.

[69] William Chan et al. “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition”. In: *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2016, pp. 4960–4964.

[70] Alex Graves. “Sequence transduction with recurrent neural networks”. In: *arXiv preprint arXiv:1211.3711* (2012).

[71] Alex Graves et al. “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks”. In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 369–376.

[72] Jan K Chorowski et al. “Attention-based models for speech recognition”. In: *Advances in neural information processing systems* 28 (2015).

[73] Chao Zhang et al. “Improving the fusion of acoustic and text representations in RNN-T”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 8117–8121.

[74] Shinji Watanabe et al. “ESPnet: End-to-end speech processing toolkit”. In: *arXiv preprint arXiv:1804.00015* (2018).

[75] Alec Radford et al. “Robust speech recognition via large-scale weak supervision”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 28492–28518.

[76] Alexei Baevski et al. “wav2vec 2.0: A framework for self-supervised learning of speech representations”. In: *Advances in neural information processing systems* 33 (2020), pp. 12449–12460.

[77] Awni Hannun et al. “Deep speech: Scaling up end-to-end speech recognition”. In: *arXiv preprint arXiv:1412.5567* (2014).

[78] Navdeep Jaitly and Geoffrey E Hinton. “Vocal tract length perturbation (VTLP) improves speech recognition”. In: *Proc. ICML workshop on deep learning for audio, speech and language*. Vol. 117. 2013, p. 21.

[79] Alec Radford et al. *Whisper*. [Online; accessed 2025]. 2022. URL: `https://github.com/openai/whisper`.

[80] Ashish Seth et al. “Stable distillation: Regularizing continued pre-training for low-resource automatic speech recognition”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 10821–10825.

[81] Thomas Rolland and Ana Abad. “Introduction to Partial Fine-tuning: A Comprehensive Evaluation of End-to-End Children’s Automatic Speech Recognition Adaptation”. In: *Interspeech 2024*. 2024. URL: `https://www.isca-archive.org/interspeech_2024/rolland24_interspeech.pdf`.

[82] B. Sertolli et al. “Representation Transfer Learning from Deep End-to-End Speech Recognition Networks for the Classification of Health States from Speech”. In: *Computer Speech Language* 68 (2021), p. 101204. DOI: `10.1016/j.csl.2021.101204`. URL: `https://doi.org/10.1016/j.csl.2021.101204`.

[83] Laurence Gillick and Stephen J Cox. “Some statistical issues in the comparison of speech recognition algorithms”. In: *International Conference on Acoustics, Speech, and Signal Processing,* IEEE. 1989, pp. 532–535.

[84] Naeem Talha. *WER Statistical Significance Test*. [Online; accessed 2025]. 2017. URL: `https://github.com/talhanai/wer-sigtest`.

[85] Nelleke Oostdijk et al. “The Spoken Dutch Corpus. Overview and First Evaluation.” In: *LREC*. Athens, Greece. 2000, pp. 887–894.

[86] David A van Leeuwen et al. “Results of the N-Best 2008 Dutch speech recognition evaluation”. In: (2009).

[87] Catia Cucchiarini et al. “Jasmin-cgn: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality”. In: (2006).

[88] Anmol Gulati et al. “Conformer: Convolution-augmented transformer for speech recognition”. In: *arXiv preprint arXiv:2005.08100* (2020).

[89] Yuanyuan Zhang et al. “Improving child speech recognition with augmented child-like speech”. In: *arXiv preprint arXiv:2406.10284* (2024).

[90] Jenthe Thienpondt and Kris Demuynck. “Transfer Learning for Robust Low-Resource Children’s Speech ASR with Transformers and Source-Filter Warping”. In: *arXiv preprint arXiv:2206.09396* (2022).

[91] Sanchit Gandhi. *Fine-Tune Whisper For Multilingual ASR with Transformers*. URL: `https://huggingface.co/blog/fine-tune-whisper`.

[92] Edward Ma. *NLP Augmentation*. https://github.com/makcedward/nlpaug. 2019.

[93] S Shahnawazuddin et al. “Exploring the Role of Speaking-Rate Adaptation on Children’s Speech Recognition”. In: *2018 International Conference on Signal Processing and Communications (SPCOM)*. IEEE. 2018, pp. 21–25.

[94] Daniel S Park et al. “Specaugment: A simple data augmentation method for automatic speech recognition”. In: *arXiv preprint arXiv:1904.08779* (2019).

[95] Hemant Kumar Kathania et al. “Explicit pitch mapping for improved children’s speech recognition”. In: *Circuits, Systems, and Signal Processing* 37 (2018), pp. 2021–2044.

[96] Chris Donahue, Bo Li, and Rohit Prabhavalkar. “Exploring speech enhancement with generative adversarial networks for robust speech recognition”. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 5024–5028.

[97] Lars Rumberg et al. “Age-Invariant Training for End-to-End Child Speech Recognition Using Adversarial Multi-Task Learning.” In: *Interspeech*. 2021, pp. 3850–3854.