university of groningen

campus fryslân

Fine-Tuning Whisper for Dutch-Speaking Autistic Children: Adapting ASR to Atypical Speech in Low-Resource Settings

Hantao Yu

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Fine-Tuning Whisper for Dutch-Speaking Autistic Children: Adapting ASR to Atypical Speech in Low-Resource Settings

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Xiyuan Gao (Voice Technology, University of Groningen)
with the second reader being
Dr. Shekhar Nayak (Voice Technology, University of Groningen)

Hantao Yu (S5910587)

June 11, 2025

Acknowledgements

First, I would like to express my heartfelt gratitude to my supervisor, Gao Xiyuan, for her generous support and patient guidance throughout the thesis process. Her dedication, responsibility and attention to detail were truly encouraging, and I have learned a lot from his feedback and supervision.

My sincere thanks also go to Dr. Matt Coler, director of the Voice Technology programme. I deeply appreciate his efforts in creating and leading this unique programme, which gave me the opportunity to spend a fulfilling and meaningful year in this field.

I am also grateful to all the teaching members in the Voice Technology team—Joshua, Vass, Phat, Shekhar and Matt, for their inspiring courses and continued support during the programme.

To all my classmates, thank you for sharing this journey together. Your support and companionship made this intense year both manageable and memorable.

Finally, I want to thank myself. I never thought I could enter a field that was entirely new to me, filled with challenges I had never encountered before, and make it through, learning so much along the way. And to the unnamed plant on my windowsill.

Abstract

Children with Autism Spectrum Disorder (ASD) often exhibit atypical prosody and disfluency patterns, posing challenges for automatic speech recognition (ASR) systems. While large-scale models like Whisper have achieved strong general performance, their effectiveness on neurodivergent speech in low-resource languages remains underexplored. This study focuses on Dutch, a relatively underrepresented language in ASD ASR research, and investigates how task-specific fine-tuning of the Whisper-medium model can improve recognition of Dutch speech from autistic children. The main experiment involves baseline fine-tuning across seven speaker group combinations (TD (typical developing children), ADHD, ASD, and their mixes). And the study is complemented by exploratory experiments using parameter-efficient LoRA fine-tuning.

Results show that fine-tuning significantly improves recognition performance, particularly when ASD speech is included in training. The best baseline configuration (TD+ASD+ADHD) reduced Word Error Rate (WER) from 43.12% (zero-shot) to 26.43%, while LoRA fine-tuning with ASD-only data further reduced WER to 23.20%, underscoring the impact of prosody-aligned training even under low-resource constraints. Error analysis revealed reductions in deletion and substitution errors, and better recognition of disfluencies such as fillers and repetitions. Statistical tests (e.g., Mann-Whitney U) confirmed the significance of performance differences across training conditions ($p < 0.05$), favoring ASD-inclusive models.

These findings emphasize the importance of prosodic alignment and domain relevance in adapting ASR systems for neurodivergent speakers. This work contributes both methodologically, by comparing full and parameter-efficient fine-tuning strategies, and practically, by advancing inclusive speech recognition solutions in low-resource, underserved populations.

Keywords: Speech Recognition, Whisper, Fine-Tuning, Atypical Speech, Autism Spectrum Disorder (ASD), Child Speech, Prosody, Disfluency

Contents

1 Introduction

1.1 ASR Development and Its Relation to Prosody and Fluency

Speech is one of the most fundamental and intuitive forms of human communication (Kohler, 2017). Its naturalness and efficiency make it a preferred mode of interaction, not only in interpersonal contexts but also in modern Human-Computer Interaction (HCI). In this context, Automatic Speech Recognition (ASR) plays a central role by enabling machines to understand and process spoken language.

Since the 1950s, ASR technology has evolved through several major stages. Early systems relied on simple pattern-matching techniques, while later approaches used statistical models such as Hidden Markov Models (HMMs). In recent decades, the rise of neural networks has led to a breakthrough in recognition performance, with models like Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs) becoming mainstream. Today, most cutting-edge ASR systems adopt an End-to-End (E2E) architecture, often based on Transformer models. OpenAI's Whisper is one such system, known for its high performance across multiple languages and environments (Malik et al., 2021; Prabhavalkar et al., 2024).

Beyond simple transcription, ASR is also embedded in broader applications such as voice assistants, emotion detection, and speaker identification. In these contexts, it is not enough for ASR to capture just the words spoken; it must also interpret prosodic and paralinguistic cues, such as intonation, rhythm, and emphasis, that convey meaning and speaker intent.

Two important aspects of speech that affect ASR performance are prosody and fluency. Prosody refers to the suprasegmental features of speech, such as pitch (fundamental frequency), duration, loudness (intensity), and rhythm, that shape the overall intonation and phrasing of spoken language. These cues help ASR systems in detecting sentence boundaries, inserting appropriate punctuation, and resolving ambiguities in meaning (Vicsi & Szaszák, 2010). Fluency, on the other hand, relates to the smoothness and continuity of speech. Fluent speech typically contains fewer pauses, repetitions, and interruptions, allowing ASR systems to map acoustic signals to words more reliably, especially in spontaneous or noisy conditions (Bhardwaj et al., 2021).

1.2 ASD Children's Speech and ASR Challenges

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that affects how individuals interact, behave, and communicate (C et al., 2018). Children's speech in general differs from adult speech in terms of how sounds are produced (articulation), how they are heard (acoustics), and how language is used. For children with ASD, these differences are even more pronounced. Their speech often shows atypical prosody, for example, using unusually high or flat pitch, speaking too slowly or too quickly, or placing stress on the wrong words. They also tend to speak with disfluencies, including frequent fillers like "um" or "uh", repeating words, or stopping mid-sentence and restarting (Asghari et al., 2021). These speech patterns are found across many languages, including Dutch.

These atypical features make it harder not only for people to understand what the child is saying,

but also for ASR systems to transcribe it accurately. Most ASR systems, such as OpenAI's Whisper, are trained on clear and fluent adult speech (D. Mujtaba et al., 2024). As a result, they often struggle with speech that is disfluent or atypical. For instance, Whisper tends to ignore or smooth out elements like pauses, repetitions, or fillers (e.g., "um", "uh"), treating them as if they weren't there. This can lead to incorrect or incomplete transcriptions. These problems become even more severe in natural, conversational speech, where people speak in a less organized way and the boundaries between phrases or sentences are harder to detect (Graham & Roll, 2024).

1.3 Research Gap and Purpose

Although many studies have documented that children with ASD often exhibit unusual prosodic features, such as abnormal pitch or timing, few have examined how these features directly contribute to recognition errors in ASR systems (Asghari et al., 2021). Moreover, existing ASR studies on ASD or disfluent speech have primarily focused on English, leaving non-English languages, such as Dutch, largely underexplored, despite Dutch being a widely spoken language with its own unique prosodic patterns (Fuckner et al., 2023). In addition, there is a lack of research on adapting Whisper, the widely used end-to-end ASR model, to the speech of neurodiverse children. In particular, few efforts have been made to fine-tune Whisper using small amounts of spontaneous, disfluent speech, which more accurately reflects real-world communication challenges.

To address these gaps, this study proposes a targeted fine-tuning of the Whisper model using under two hours of Dutch speech data from three groups of children: typically developing (TD), children with ADHD, and children with ASD. The aim is twofold: first, to improve Whisper's ability to accurately transcribe speech from children with ASD; second, to examine how specific prosodic and disfluency patterns in their speech are related to the types of recognition errors made by the model.

1.4 Research Questions

Although Dutch is not considered a low-resource language, current ASR systems, including advanced end-to-end models like Whisper, still struggle with underrepresented speech types. These include speech from speakers with atypical developmental profiles, such as children with Autism. The unique acoustic and linguistic characteristics of ASD speech are typically not well captured by large-scale training datasets. This mismatch often leads to higher word error rates (WER) for this population.

This study addresses the following two core research questions:

> **Can fine-tuning the Whisper model on a small dataset of Dutch child speech (including TD, ADHD, and ASD) improve ASR performance for ASD speech, as measured by WER?**
>
> **What prosodic and disfluency-related features characterize the speech of Dutch-speaking autistic children, and how are these features related to recognition errors produced by Whisper?**

No existing study has specifically investigated how to improve ASR for Dutch-speaking autistic children. Furthermore, studies examining the prosodic and disfluent characteristics of this population remain limited. Based on previous literature involving other languages and clinical speech domains, this study proposes the following hypotheses:

> **Fine-tuning Whisper on a combination of Dutch child speech from ASD, TD, and ADHD groups will result in a measurable improvement in WER, potentially around 10%, as observed in similar low-resource or atypical speech settings (Gale et al., 2019).**
>
> **Recognition errors will show significant associations with specific prosodic (e.g., pitch variability, utterance duration) and disfluency-related features (e.g., repetitions, filler words), consistent with prior findings in clinical and child speech studies (Soleymanpour et al., 2022; Tobin et al., 2024).**

To explore these questions, the study conducts a series of fine-tuning experiments using the Whisper-medium model. Speech data from TD, ADHD, and ASD children are used separately and in combination to evaluate their impact on recognition accuracy and error types. The goal was to evaluate how fine-tuning with data from different populations and their combinations affects the Whisper model's performance on ASD speech. Subsequently, I conducted an error analysis to investigate the types and patterns of recognition errors.

2 Literature Review

To position this study within the broader research landscape, the literature review part synthesizes prior findings on prosodic and disfluency patterns in children with ASD, the limitations of general ASR models on non-typical speech, and recent strategies for domain-specific fine-tuning. This review not only contextualizes the challenges this thesis addresses but also clarifies how each technical barrie, such as prosodic mismatch, model generalization, or data scarcity, can shape the methodological decisions made in this work.

2.1 Prosodic and Disfluency Features in ASD Children's Speech

Prosody and disfluency are critical elements of spoken language that extend beyond lexical content. Prosody refers to the suprasegmental features of speech, such as pitch (fundamental frequency, F0), duration, intensity, and rhythm. These cues help listeners understand sentence structure, identify which words are emphasized, and interpret the speaker's intent, for example, whether they are asking a question or making a statement (Asghari et al., 2021). Disfluency refers to interruptions or irregularities in speech, such as filler words like "uh" or "um", repeated words, stretched sounds, and self-corrections. In typical speakers, these features help manage turn-taking or signal hesitation during conversations (Zorić, 2024).

Importantly, both prosodic and disfluency phenomena are not only critical for human comprehension, but also represent potential risk factors for errors in automated speech technologies. Speech patterns that diverge from neurotypical norms, such as irregular pauses or exaggerated pitch contours, can disrupt the acoustic patterns expected by ASR models, leading to reduced transcription accuracy.

2.1.1 Atypical Prosody in Children with ASD

Children with Autism Spectrum Disorder (ASD) frequently exhibit atypical prosodic features. These include unusual pitch patterns (such as differences in contour, range, or variability), abnormal duration, atypical intensity, and misplaced stress. These prosodic deviations are not only perceptually salient but also objectively measurable using acoustic tools such as Praat. They have been shown to impact both speech intelligibility and how listeners perceive the speaker (Asghari et al., 2021).

However, empirical findings on prosodic differences between ASD and typically developing (TD) children have been inconsistent. For example, some studies report that individuals with ASD exhibit a wider pitch range than TD peers (Diehl et al., 2009; Lehnert-LeHouillier et al., 2020), while others have found a narrower range (Santen et al., 2010) or no significant difference at all (Paul et al., 2008). The inconsistency of results not only complicates theoretical understanding but also limits the application of such findings in speech technology and clinical screening tools.

One likely explanation for these inconsistencies is the wide variation in data collection protocols. For instance, some studies use spontaneous conversation or open-ended storytelling tasks, which tend to elicit more natural prosodic variation. Others rely on structured tasks like reading scripted materials, which may suppress individual prosodic traits (Godin & Hansen, 2015). Such method-

ological differences directly influence prosodic outcomes, including pitch, duration, and intensity. This lack of methodological standardization undermines the reliability of cross-study comparisons and complicates efforts to generalize findings to real-world speech contexts.

In terms of duration, some studies have reported significant differences between ASD and TD speakers (Diehl & Paul, 2013; Filipe et al., 2014). Others highlight the role of emotional context in shaping duration in TD children, a factor that has yet to be systematically investigated in ASD populations (Hubbard et al., 2017). Regarding intensity, most research finds no substantial group differences (Diehl & Paul, 2013). However, Olivati et al. (2017) observed significantly higher maximum and minimum intensity levels in the speech of children with ASD. Across all prosodic dimensions, individual variability may also be influenced by factors such as age, gender, IQ, and expressive language ability. These participant-level differences often vary within and between diagnostic groups, making it difficult to generalize findings.

Given these sources of variability, both methodological and participant-related, it remains challenging to draw consistent conclusions about prosodic patterns in ASD speech. Additionally, most studies focus on isolated prosodic features without modeling their interactional or contextual dynamics, which limits their ecological validity and applicability in dynamic ASR environments. Moreover, many studies rely on experimental designs that lack ecological validity and do not include systematic acoustic labeling. As a result, the implications of atypical prosody for downstream technologies such as ASR remain underexplored.

2.1.2 Disfluency in ASD Speech

While prosodic features in ASD speech show considerable variability across individuals and studies, some speech characteristics, particularly disfluencies, emerge more consistently and prominently. A frequently observed feature is the presence of disfluencies, such as repetitions, filled pauses, and irregular pauses, that occur more often and in less predictable positions than in neurotypical speech (Lake et al., 2011). These disfluencies often appear at syntactically or semantically inappropriate points, making the speaker's intended meaning harder to understand (Zorić, 2024). In neurotypical speech, disfluencies often serve pragmatic functions, such as signaling hesitation, corrections, or managing turn-taking. In contrast, disfluencies in ASD speech frequently lack such contextual relevance and can actively hinder listener interpretation.

These non-pragmatic disfluencies tend to occur in semantically unrelated or socially inappropriate places, which can confuse listeners and reduce overall communicative effectiveness. Nevertheless, many studies stop at descriptive analysis, without developing predictive or computational models that can capture such disfluency behaviors in practical applications. The occurrence and form of disfluencies may vary with factors such as age, language environment, and speaking task (Holler & Levinson, 2019). Moreover, much of this research is based on English-speaking populations, with limited attention to language-specific or cross-linguistic prosodic norms, raising questions about generalizability. To systematically capture this variation, Lake et al. (2011) proposed a classification scheme that divides disfluencies in ASD speech into three main types: stalling, repair, and fragmentation. Among these, stalling and fragmentation are especially common in ASD speech, particularly during spontaneous narratives that place high demands on planning and fluency.

In addition to disfluency patterns, prosodic irregularities in ASD speech also show cross-linguistic consistency. A meta-analysis found that features such as flattened pitch contours, reduced pitch range, and irregular timing were consistently observed across English, Dutch, and French (Asghari et al., 2021). This suggests that these atypical prosodic cues may reflect broader cognitive-perceptual traits, rather than being language-specific phenomena.

2.1.3 Prosodic Features in ADHD Speech

While much of the existing literature has focused on the prosodic and disfluency patterns of ASD, emerging research suggests that children with Attention-Deficit/Hyperactivity Disorder (ADHD) may also exhibit distinct prosodic abnormalities. Although the underlying neurodevelopmental mechanisms differ, both populations display atypical speech profiles that challenge conventional models of prosody and fluency development. In contrast to the often flattened or rigid prosodic contours observed in ASD, children with ADHD tend to produce exaggerated and unstable pitch and intensity variations. For instance, Cassol-Jr et al. (2010) found that during spontaneous mother-child interactions, children with ADHD exhibited highly variable loudness and pitch, resulting in an erratic paralinguistic style. These atypical prosodic patterns may reflect underlying deficits in executive function, particularly in inhibitory control, which disrupt the real-time regulation of vocal output.

Although research on ADHD speech is more limited, the available findings suggest a distinct trajectory. Whereas prosodic anomalies in ASD are often linked to social-pragmatic impairments, those in ADHD appear more related to impulsivity and reduced control over vocal modulation. This distinction highlights the importance of developing disorder-specific models for analyzing and interpreting speech characteristics, especially in clinical and technological contexts such as ASR.

Despite increasing interest in atypical prosody, the literature remains fragmented. In the case of ASD, many studies rely heavily on perceptual judgments rather than systematic acoustic analysis, limiting cross-study comparability and integration into computational models. For ADHD, research is even more sparse and often lacks the methodological rigor needed for reliable cross-diagnostic comparison.

Furthermore, spontaneous speech corpora that capture children's natural vocal behavior remain critically scarce, especially for underrepresented languages such as Dutch. Few publicly available datasets, such as ASDBank, provide the level of annotation and linguistic diversity needed to support robust acoustic modeling or to train inclusive ASR systems. This highlights the urgent need for large-scale, systematically annotated corpora to advance theunderstanding of prosody and disfluency in neurodivergent child populations. Even when such corpora exist, their annotations are often inconsistent or lack standard prosodic labeling, making it difficult to train generalizable or explainable ASR models.

2.2 Challenges of ASR on Child and Atypical Speech

ASR systems consistently perform worse on children's speech due to notable acoustic and articulatory mismatches with adult-trained models. Compared to adults, children typically produce speech with higher fundamental frequency (F0), shorter vocal tract lengths, less stable articulation, and greater variability in speaking rate (Sobti et al., 2024). These physiological and developmental differences substantially elevate WER, particularly in spontaneous, unscripted speech.

For children with ASD, these baseline challenges are compounded by atypical prosodic cues and context-insensitive disfluencies, as discussed in Section 2.1. Features such as irregular pausing, misplaced stress, and filled pauses disrupt standard segmentation cues and acoustic-to-linguistic mappings, leading to increased substitution, insertion, and deletion errors. These problems are particularly acute in spontaneous narrative tasks, where such disfluencies occur unpredictably and interfere with both syntactic alignment and semantic coherence.

Despite the importance of these issues, there remains a striking lack of controlled experimental studies that systematically compare ASR performance between ASD and typically developing children. This gap is especially concerning given the increasing reliance on ASR in educational, clinical, and assistive settings. The absence of such comparisons not only limits theunderstanding of recognition disparities but also impedes the development of inclusive and generalizable speech technologies.

2.2.1 Whisper ASR Model: Architecture and Bias

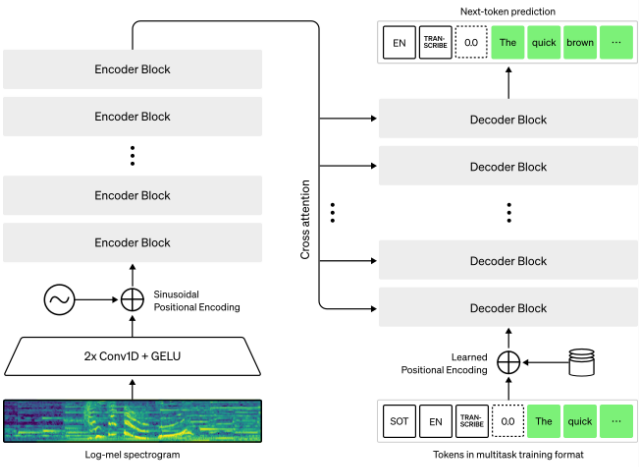


Figure 1: Whisper ASR Architecture

Whisper (Radford et al., 2022) is a large-scale transformer-based ASR model developed by OpenAI, designed to handle speech recognition, translation, and language identification through a unified encoder-decoder architecture. Its sequence-to-sequence framework, unlike traditional CTC-based systems, allows Whisper to model long-range dependencies and contextual semantics across multilingual domains. Moreover, Whisper supports transcription in nearly 100 languages and is pretrained

on over 680,000 hours of speech, making it highly adaptable to low-resource and cross-lingual settings. Its architecture is as shown in Figure 1.

Despite its architectural sophistication, Whisper is trained on weakly supervised data sourced primarily from online platforms such as YouTube and broadcast media. Critically, this dataset overwhelmingly consists of normative, fluent, adult speech. Speech from children, accented speakers, disfluent speakers, or neurodivergent individuals is severely underrepresented. This data imbalance introduces systematic bias into the model's learned representations.

As a result, Whisper exhibits notable performance degradation when confronted with marginalized speech populations, including children, L2 speakers, and individuals with dysarthria or neurodevelopmental conditions (Fuckner et al., 2023; Jain, 2023). In particular, when processing ASD speech, Whisper has been observed to hallucinate content, omit filled pauses, or restructure disfluent segments to resemble grammatical fluency (Mujtaba et al., 2024). These distortions reflect not just a recognition error but an implicit prioritization of fluency over fidelity, thereby misrepresenting speakers who fall outside normative expectations.

Nonetheless, Whisper's strong zero-shot performance and multilingual flexibility make it an attractive candidate for fine-tuning in atypical speech domains. In recent benchmarking on Dutch ASR tasks, Whisper demonstrated top-tier performance on standard datasets such as CGN and JASMIN (as reported by the ASR-NL initiative https://github.com/opensource-spraakherkenning-nl), outperforming many traditional Kaldi-based or wav2vec2 systems on clean read speech. While these benchmarks primarily involve adult and scripted speech, they suggest Whisper's capacity to generalize well to Dutch, especially when adapted with even modest amounts of in-domain data.

Furthermore, although some studies have tested Whisper on child speech, they predominantly rely on scripted or read-aloud tasks, which fail to capture the naturalistic prosodic variation and irregularity found in spontaneous dialogue. This methodological gap obscures the true extent of Whisper's limitations in real-world conditions, particularly for underrepresented languages such as Dutch and underexplored groups such as children with ASD or ADHD.

These limitations are not merely technical oversights but structural barriers that reinforce exclusion. Whisper's training paradigm and decoding biases collectively restrict its usability for inclusive speech technologies. However, its success in prior Dutch-language adaptation studies and its flexible architecture make it well-suited for domain-specific fine-tuning. The need for targeted, inclusive ASR modeling, especially on spontaneous, neurodivergent child speech in low-resource languages, thus presents both a practical opportunity and an ethical imperative.

2.2.2 Prosody-ASR Error Correlations

A growing body of empirical research has revealed that prosodic features, including pitch variability, segmental duration, and pausing behavior, strongly influence the accuracy of Automatic Speech Recognition (ASR) systems, particularly in challenging domains such as spontaneous or child-directed speech. For instance, Goldwater et al. (2010) demonstrated that short duration and reduced prosodic prominence significantly increased word error rates in ASR transcriptions. Simi-

larly, Stoyanchev et al. (2012) showed that prosodic anomalies, including irregular pause placement and extreme pitch shifts, were strongly predictive of recognition errors. Their prosody-based error prediction models achieved AUC values up to 0.84, highlighting the diagnostic value of suprasegmental speech features in identifying high-risk regions for ASR failure.

Recent work has shown that prosodic information can be effectively integrated into self-supervised speech models to improve recognition accuracy. For example, ProsodyBERT, a model that embeds prosodic cues like pitch and rhythm into the Transformer architecture, achieving substantial reductions in substitution and deletion error rates (Y. Hu et al., 2022). These results underscore that temporal and tonal dynamics are integral, not peripheral, components of the speech signal, which conventional ASR models like Whisper may inadequately model.

However, most of these studies focus on standard child corpora and high-resource languages, and few explicitly examine how prosodic irregularities interact with neurodivergence or language-specific constraints. Moreover, while models like HuBERT have incorporated prosody through architectural adjustments (Hsu et al., 2021), Whisper has yet to systematically integrate or leverage such information. This omission may partially explain its underperformance on spontaneous, disfluent speech.

Taken together, these studies offer compelling motivation to investigate the role of prosody in ASR error patterns. Yet, a critical research gap remains: there is still little empirical evidence linking prosodic features with ASR recognition errors in neurodivergent children's speech, particularly in underrepresented languages like Dutch. This gap directly motivates theproposed correlation analysis between prosody/disfluency patterns and Whisper misrecognitions in Dutch-speaking children with ASD.

2.3 Adapting ASR to Low-Resource and Atypical Speech Domains

2.3.1 Fine-Tuning Whisper for Low-Resource and Atypical Speech

Fine-tuning refers to the adaptation of a pre-trained model to a specialized task using a smaller, domain-specific dataset. In the case of Whisper, originally trained on over 680,000 hours of predominantly normative, adult, English-language speech, fine-tuning allows the model to better accommodate underrepresented forms of speech, including spontaneous, disfluent, or child speech from marginalized populations.

This approach proves especially beneficial in low-resource settings, where collecting large-scale annotated speech corpora is infeasible. Recent work has demonstrated that Whisper can achieve significant gains with very limited training data. For example, Rijal et al. (2024) fine-tuned Whisper on just 80 minutes of Nepali speech and achieved a relative WER reduction of 35.6% compared to the baseline. Similarly, Ghimire et al. (2024) reported that parameter-efficient fine-tuning of Whisper using only 100 minutes of Nepali speech yielded WER improvements exceeding 30% over zero-shot performance, highlighting Whisper's strong adaptation potential in low-resource environments.

However, while these results are promising, they are often limited to scripted, monolingual, or relatively clean speech settings. Their generalizability to spontaneous, disfluent, and neurodivergent

speech remains largely untested. This is especially problematic given that the populations most in need of inclusive ASR solutions are typically those furthest from the normative training distributions of large models.

To address overfitting risks and computational constraints in such data-scarce scenarios, recent studies have adopted parameter-efficient fine-tuning methods. Techniques like partial fine-tuning (which freezes most model layers), LoRA (which adds low-rank trainable matrices), and BAFT offer lightweight alternatives to full fine-tuning (Bhardwaj et al., 2022; Liu et al., 2024; Mujtaba et al., 2024). These approaches lower the barrier for adapting Whisper to specific speech domains, but their performance on disfluent child speech, especially in typologically distinct or underrepresented languages, remains underexplored.

Despite technical advances in adaptation strategies, there is still a striking lack of research applying Whisper to spontaneous speech from neurodivergent children in languages such as Dutch. No existing studies, to theknowledge, have systematically evaluated how fine-tuning Whisper with Dutch ASD speech affects model performance or fairness. This gap is not merely empirical, it signals a broader neglect of linguistic diversity and neurodivergent populations in speech technology research. The study addresses this need by evaluating the efficacy of low-resource Whisper fine-tuning on the Dutch ASDBank corpus.

2.3.2 Parameter-Efficient Fine-Tuning: LoRA and Alternatives

In low-resource domains such as disfluent or neurodivergent speech modeling, full fine-tuning of large-scale ASR models like Whisper is often computationally prohibitive and susceptible to overfitting. To address this, the field has increasingly turned to Parameter-Efficient Fine-Tuning (PEFT) techniques, which enable task-specific adaptation by training only a small fraction of the model's parameters (Ashvin et al., 2024; Müller-Eberstein et al., 2024; Zhang et al., 2025).

Among these, Low-Rank Adaptation (LoRA) has emerged as a leading method. Rather than updating the full parameter space, LoRA inserts pairs of low-rank trainable matrices into the attention layers while keeping the original weights frozen (E. J. Hu et al., 2021). This dramatically reduces memory usage and computational cost, while still allowing the model to adapt effectively to new domains.

Alternative PEFT approaches have also been explored:

- Adapter Tuning adds small bottleneck layers between transformer blocks, preserving the main architecture but increasing memory and latency costs across all layers;
- BitFit, which updates only bias terms, offers minimal training overhead but struggles with complex, acoustically variable tasks;
- Prefix Tuning, effective in text-based transformers, prepends learnable prompts to input embeddings, but its reliance on token-level semantics limits its utility for speech models where timing and prosody matter more than discrete tokens.

While LoRA's utility is well-established in NLP, its application in speech modeling, particularly for Whisper, remains relatively unexplored. Only a handful of studies (e.g., Liu et al. (2024); Mujtaba et al. (2024)) have tested LoRA on ASR models, and these primarily target clean or scripted datasets. The feasibility and impact of LoRA on atypical or disfluent speech, especially in low-resource and non-English contexts, remains largely an open question.

Moreover, Whisper's unique encoder-decoder structure poses specific challenges for PEFT. Unlike uni-directional encoders in most language models, Whisper processes temporal acoustic patterns and long-span dependencies jointly, raising questions about where and how LoRA modules should be inserted to maximize efficacy. Existing PEFT benchmarks seldom address these speech-specific architectural nuances, creating a gap in both theoretical understanding and empirical validation.

Given the extremely limited availability of annotated Dutch speech from children with ASD or ADHD, and the presence of disfluency patterns that diverge sharply from Whisper's normative training data, LoRA offers a promising yet under-tested pathway for efficient model adaptation. By incorporating LoRA into our fine-tuning of Whisper, this study aims to extend the applicability of PEFT methods to spontaneous, neurodivergent child speech, an area that has been critically underrepresented in both ASR research and model evaluation frameworks.

2.4 Summary and Identified Gaps

The reviewed literature underscores the complex interplay between prosody, disfluency, and ASR performance in neurodivergent child speech. Studies on ASD and ADHD speech have identified distinct suprasegmental patterns, ranging from flat intonation in ASD to exaggerated pitch variability in ADHD, that interfere with standard acoustic-linguistic mappings. These deviations are further compounded by irregular pausing and filled pauses, which distort the timing cues relied upon by ASR models. Despite this, the majority of prior work has focused on adult or TD populations, leaving neurodivergent child speech significantly underexplored.

Although state-of-the-art ASR models like Whisper have demonstrated robust generalization across many domains, their training data remains heavily skewed toward fluent, adult speech. This imbalance results in marked performance degradation when applied to spontaneous, disfluent, or developmentally atypical speech. Furthermore, most evaluations of Whisper's performance on child speech rely on read or scripted corpora, which do not reflect the prosodic irregularities or disfluency patterns characteristic of real-world child communication, particularly in low-resource languages like Dutch.

While prosodic features have shown promise as predictors of ASR errors, existing models have seldom incorporated them explicitly, especially within the Whisper framework. Research integrating prosodic embeddings into ASR architectures has mostly focused on standard child corpora or high-resource languages, and few studies have investigated how these features correlate with recognition errors in neurodivergent speech. Additionally, although parameter-efficient fine-tuning methods such as LoRA have shown success in natural language processing and adult ASR tasks, their application to spontaneous, disfluent child speech remains largely untested.

Finally, a critical infrastructural limitation persists: the lack of systematically annotated corpora for neurodivergent child speech in underrepresented languages. Even when such datasets exist, they often lack consistent prosodic labeling or robust disfluency annotation, hindering the development of explainable and generalizable ASR systems. In sum, despite growing interest in inclusive speech technology, three interrelated gaps remain unresolved:

Limited fine-tuning research on neurodivergent child speech, particularly in Dutch and involving spontaneous, disfluent utterances;

Lack of empirical analysis linking prosodic and disfluency features with ASR recognition errors in ASD and ADHD speech;

Underexplored use of parameter-efficient adaptation methods (e.g., LoRA) for addressing atypical speech in low-resource settings.

These gaps directly motivate the present study, which seeks to address both the technical and linguistic challenges in modeling ASR for Dutch-speaking neurodivergent children.

3 Methodology

This section outlines the methodological framework adopted to investigate the effectiveness of fine-tuning Whisper models for improving ASR performance on speech from Dutch-speaking children (TD, ADHD, ASD). Building upon the findings discussed in the literature review, which highlighted the limitations of existing ASR systems in handling non-standard speech patterns, particularly in low-resource or disordered speech contexts, this study adopts a transfer learning approach using the Whisper architecture.

The methodology consists of four main components: construction and preprocessing of a representative dataset, selection and description of the pre-trained Whisper model, design and implementation of the fine-tuning experiments across different training conditions, and quantitative evaluation of ASR performance.

Through this methodological design, the study aims to systematically examine how model architecture, training data composition, and fine-tuning strategies interact to influence ASR outcomes. The following subsections provide a detailed account of each component. The complete code and experimental setup used in this thesis are available at github[1].

3.1 Dataset Description

3.1.1 Corpus Overview

The speech data used in this study were drawn from the ASDBank Asymmetries Corpus (Kuijper et al., 2015), a linguistically rich Dutch corpus designed to facilitate comparative analysis across neurodevelopmental profiles. The corpus includes recordings of 86 children aged 6-12, categorized into three diagnostic groups: Typically Developing (TD), Attention-Deficit/Hyperactivity Disorder (ADHD), and Autism Spectrum Disorder (ASD).

Each child was recorded producing four picture-based narrative retellings (Pirate, Ballerina, Princess, and Indian), following the CHAT transcription protocol. The elicitation setting was semi-spontaneous and child-friendly, yielding approximately 1 to 2 minutes of speech per speaker. ASD diagnoses were made based on at least one gold-standard instrument, the Autism Diagnostic Interview-Revised (ADI-R) or the Autism Diagnostic Observation Schedule (ADOS). Notably, around 60% of ASD participants met criteria on both instruments, while the remainder satisfied only one but exhibited pronounced social-communicative features consistent with ASD. This inclusion strategy enhances not only the diagnostic validity but also the ecological validity of the dataset, aligning with contemporary trends in autism research that emphasize functional diversity and real-world communicative variation.

To ensure acoustic quality and modeling reliability, additional filtering was performed on the original recordings. Speech segments with overlapping speech, excessive background noise, or degraded audio fidelity were excluded. The data that is used in this study is as shown in table 1:

[1] https://github.com/maxyuht/VT_thesis.git

Table 1: Dataset Information

Group	**Number of Speakers**	**Total Duration (min)**	**Average Duration per Speaker (min)**	**Approx. Average Utterance Length**
TD	30	64.65	2.16	˜6-8 seconds
ADHD	19	50.35	2.65	˜6-10 seconds
ASD	37	26.84	0.73	˜4-6 seconds
Total	86	141.84	–	–

Compared to TD and ADHD groups, the ASD group contributed significantly shorter average durations, partly due to speech production challenges typical of this population. Nevertheless, the corpus offers valuable insights into how ASR models handle speech diversity across neurodevelopmental conditions.

To evaluate the effectiveness of Whisper fine-tuning in recognizing atypical speech patterns, particularly those characteristics of ASD, a speaker-independent data split was applied to the 37 ASD participants in the curated ASDBank corpus. The split was performed randomly, with the following distribution:

- Training set: 12 ASD speakers, used for model fine-tuning
- Validation set: 10 ASD speakers, used for tuning monitoring and early stopping
- Test set: 15 ASD speakers, held out for final evaluation

All TD and ADHD speakers were included exclusively in the training set to simulate a realistic low-resource scenario where models are trained on typical or near-typical speech but must generalize to more idiosyncratic clinical speech.

This design serves two purposes: first, it ensures speaker independence across subsets, thereby preventing data leakage and inflated performance estimates; second, it enables an empirical investigation into how cross-group transfer, from TD, ADHD to ASD speech, impacts ASR performance. By reserving ASD speech solely for evaluation and validation, the study ensures that generalization metrics reflect model robustness in clinically relevant target conditions.

To systematically explore this question, the study adopted seven fine-tuning paradigms, each designed to probe different aspects of model generalization and the incremental value of ASD training data:

- TD only

- ADHD only
- ASD only
- TD+ADHD
- ADHD+ASD
- TD+ASD
- TD+ADHD+ASD

These paradigms allow for fine-grained comparison of training sources and provide insight into whether fine-tuning on more neurotypical data benefits or hinders ASR performance on atypical ASD speech, and how the inclusion of even limited ASD samples might shift recognition accuracy.

3.1.2 Preprocessing and Annotation Normalization

To prepare the audio data for model training, a standardized preprocessing pipeline was implemented. All original recordings in MP3 format were first converted to WAV format with a sampling rate of 16 kHz, mono channel, and 16-bit PCM encoding. Each speaker's speech was then manually segmented into utterances of less than 30 seconds using Praat, ensuring compatibility with Whisper's input length constraints and minimizing memory overflow during fine-tuning.

For ASR evaluation purposes, CHAT-format transcriptions were cleaned and converted to Praat-compatible .TextGrid files. To enable accurate and consistent Word Error Rate (WER) calculation, a set of normalization rules was applied:

- Fillers (e.g., uhm, eh) were retained to preserve disfluency patterns, which are critical for assessing ASR robustness to spontaneous speech.
- Partial words or interrupted phonemes (e.g., pi(raat)) were truncated to their audibly realized component (pi), ensuring alignment with what is actually spoken and avoiding artificial penalization in WER computation.
- Unclear or optional phonemes (e.g., (h)em) were regularized based on contextual cues ($\rightarrow$ hem), promoting transcription-listening consistency.
- CHAT meta-symbols and non-lexical markers (e.g., pause indicators, retracing markers) were stripped to eliminate mismatches between transcribed tokens and acoustic content, while still preserving lexical fidelity to the child's original utterances.

These adjustments aimed to strike a balance between linguistic accuracy and evaluation validity. By aligning transcriptions more closely with perceptible speech, the normalization reduces spurious mismatches during WER scoring and increases the interpretability of recognition errors in low-resource, spontaneous child speech contexts.

3.2 Model Architecture

In this study, I utilized OpenAI's Whisper architecture for Dutch automatic speech recognition. While various model sizes are available, I compared two pre-trained checkpoints, Whisper-medium and Whisper-large-v2, as baselines in a zero-shot setting.

The following table summarizes the two Whisper variants referenced in this research:

Table 2: Whisper ASR Models

Model	**Parameter Count**	**Disk Size**	**Description**
Whisper-Medium	769 million	1.5 GB	Mid-sized model balancing performance and efficiency
Whisper-Large-v2	1.55 billion	3.1 GB	Largest publicly available model with state-of-the-art ASR performance

Before fine-tuning, I conducted a preliminary evaluation of two pre-trained Whisper models, Whisper-Large-v2 and Whisper-Medium, on the ASD test set in a zero-shot setting, in order to establish a performance baseline. The Whisper-Large-v2 model achieved a Word Error Rate (WER) of 37.10%, outperforming Whisper-Medium, which obtained a WER of 43.12%. Despite the better baseline performance of large-v2, I chose Whisper-medium as the fine-tuning target, for the following three reasons:

1. Computational efficiency: Medium is faster to train and consumes significantly less GPU memory, which is essential for low-resource iterative experiments.

2. Overfitting prevention: Larger models are more prone to overfitting when fine-tuned on very limited data, which is the case for thesmall speech dataset.

3. Prior research support: Studies such as Zhang et al. (2025) has shown that smaller models tend to benefit more significantly from task-specific adaptation in extremely low-resource domains.

This choice balanced performance and practicality, enabling more controlled and replicable experiments across the six training paradigms.

3.3 Baseline Fine-Tuning of Whisper-Medium

3.3.1 Fine-Tuning Strategy

To balance generalization and task adaptation under data-scarce conditions, this study first employed a partial fine-tuning strategy. Specifically, the lower layers of the Whisper-medium encoder were frozen, while only four Transformer blocks (layers 8-11) were unfrozen during training. This decision was based on findings from prior work suggesting that upper layers in transformer-based ASR models encode more abstract linguistic and semantic information, which is more relevant for downstream adaptation (Liu et al., 2024; Yang, Zhang, Tao, Ma, & Qin, 2023).

This selective unfreezing allows the model to retain general acoustic knowledge from pre-training while learning task-specific patterns from limited target-domain data. Such strategies have also been shown to prevent overfitting in low-resource fine-tuning settings (Lee et al., 2024). The baseline fine-tuning flow is as shown in Figure 2.

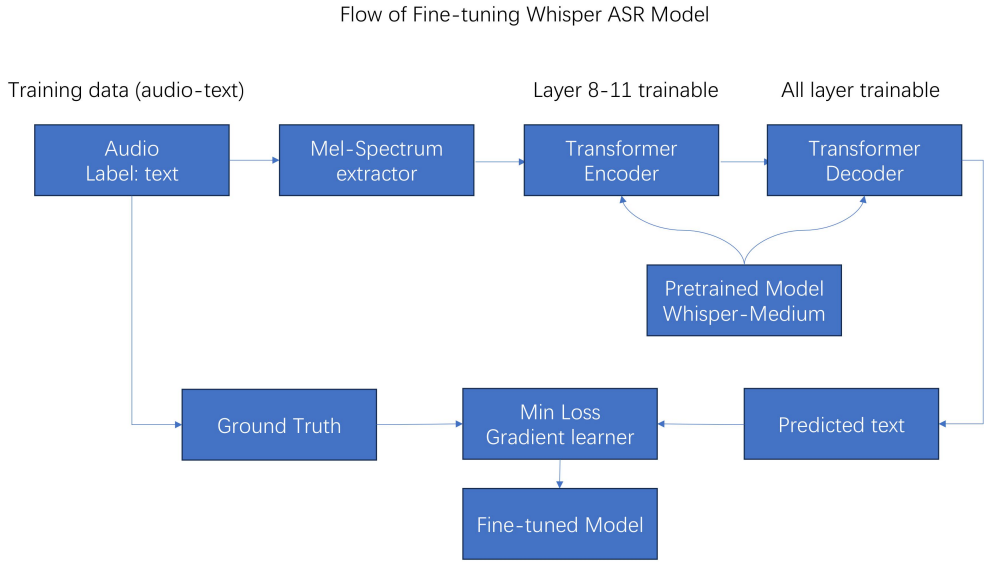


Figure 2: Whisper Baseline Fine-tuning

3.3.2 Training Configuration

Experiments were conducted on a single NVIDIA A100 GPU with 40GB of memory. Each training run used the same preprocessing pipeline to ensure consistency and reproducibility. All audio inputs were limited to 30-second segments, represented as 80-dimensional log-Mel spectrograms. The Whisper Dutch tokenizer remained frozen during training. Optimization used AdamW with a linear learning rate scheduler and warm-up.

Table 3: Training Configuration

Training Data Composition	**Max Training Steps**
TD only	100
ADHD only	100
ASD only	70
TD+ADHD	100
ADHD+ASD	100
TD+ASD	90
TD+ADHD+ASD	100

All models were trained with the following shared configuration:

- Batch size: 16
- Learning rate: 1e-5
- Warm-up steps: 5
- Evaluation interval: every 10 steps
- Optimizer: AdamW
- Scheduler: Linear decay
- Audio input: $\leq$ 30s log-Mel spectrograms
- Tokenizer: Whisper Dutch tokenizer (frozen)
- Loss: CTC loss

Each training run lasted less than 30 minutes and consumed approximately 10-12GB of GPU memory, making partial fine-tuning feasible even under limited hardware resources.

3.4 Parameter-Efficient Fine-Tuning via LoRA

3.4.1 Rationale

While partial fine-tuning helps reduce overfitting, it still requires updating a significant number of model parameters. To further reduce the computational cost and parameter footprint, this study adopted Low-Rank Adaptation (LoRA) (Hu et al., 2021), a parameter-efficient fine-tuning (PEFT) method. LoRA freezes all original model weights and introduces trainable low-rank matrices into specific attention projection layers, usually q_proj and v_proj.

This approach has shown strong performance in low-resource ASR and TTS tasks (Deng et al., 2023; Xie et al., 2023), and significantly reduces memory usage and training time. The Whisper implementation followed the publicly available Fast-Whisper-Finetuning (https://github.com/Vaibhavs10/fast-whisper-finetuning) repository, which builds on Hugging Face's peft library with Whisper-specific adjustments.

3.4.2 LoRA Configuration and Training Setup

LoRA adapters were inserted into the q_proj and v_proj layers of the Whisper encoder blocks. The rank was set to 32, with an α of 64 and dropout of 0.05. Based on the best-performing results from the previous stage, I selected two groups for LoRA adaptation: ASD only and TD+ADHD+ASD combined. Experiments were conducted on the same GPU setup as before.

Table 4: Training Configuration

Experiment Group	Max Steps	Batch Size	Grad Accum.	Learning Rate	LoRA Rank	Alpha	Dropout	Eval Step
TD + ADHD + ASD	8000	4	4	3e-6	32	64	0.05	200
ASD only	5000	2	8	3e-6	32	64	0.05	200

Shared configuration:

- Loss Function: Sequence-to-sequence cross-entropy (with built-in label smoothing)
- Audio input: $\leq$ 30s log-Mel spectrograms
- Tokenizer: Frozen Whisper Dutch tokenizer with language prefix
- Optimizer: AdamW
- Scheduler: Cosine learning rate decay with warm-up

3.5 Evaluation

To assess the performance of the fine-tuned and baseline models in recognizing speech from children with Autism Spectrum Disorder (ASD), two widely adopted metrics in automatic speech recognition (ASR) research were employed: Word Error Rate (WER) and Character Error Rate (CER). These metrics provide complementary insights into recognition quality at different linguistic levels. WER captures the overall correctness of recognized word sequences, making it a primary indicator of transcription fidelity, especially in semantic contexts. CER offers a finer-grained evaluation by accounting for character-level mismatches, which is particularly useful in short utterances or morphologically rich languages such as Dutch.

Both metrics were computed using the jiwer Python package, comparing the recognized outputs against manually curated ground truth transcriptions. Evaluation was performed on a speaker-independent test set comprising 15 autistic speakers, ensuring the robustness of results across unseen

individuals. Each fine-tuned model and zero-shot experiment are evaluated on the same test data to facilitate fair comparison.

Third, statistical tests were conducted to validate observed trends:

- Spearman's rank correlation was used to assess associations between prosodic features and WER/CER.
- OLS regression modeled WER as a function of multiple acoustic predictors.
- Mann-Whitney U tests compared prosodic patterns between high- and low-WER utterance groups.

3.6 Ethical Considerations

All data employed were pre-existing and publicly distributed under ethical research provisions. The ASDBank corpus was collected with institutional consent for linguistic analysis. This study abstains from involving any personally identifiable information or subjective evaluations, and the disfluency annotations were derived from anonymized, child-friendly scripts. All analyses and model outputs were strictly used for research purposes.

4 Results

This section presents a comprehensive evaluation of fine-tuning strategies applied to the Whisper model for improving ASR on Dutch child speech, particularly from autistic speakers. It is organized into four main parts. Section 4.1 reports baseline fine-tuning results across different combinations of training data from typically developing, ADHD, and ASD children, highlighting their impact on word and character error rates. Section 4.2 explores the effectiveness of Low-Rank Adaptation (LoRA) as a parameter-efficient alternative and compares its performance to full fine-tuning. Section 4.3 offers a text-based recognition errors and section 4.4 offers acoustic-prosodic feature correlations with ASR outcomes. This structure allows for both a quantitative performance comparison and a deeper interpretive understanding of recognition challenges in neurodivergent child speech.

4.1 Baseline Fine-Tuning Results

To evaluate the impact of different fine-tuning data configurations on ASR performance for Dutch-speaking children with autism, I conducted six baseline fine-tuning experiments using the Whisper-medium model. Each experiment involved a different subset or combination of speech from typically developing, ADHD, and ASD children. All models were evaluated on the same ASD test set using two standard metrics: Word Error Rate (WER) and Character Error Rate (CER). The table below summarizes the results.

Table 5 below summarizes the results:

Table 5: **Baseline Fine-Tuning Performance Across Speech Groups**

Fine-tuning Group	**Val-WER (%)**	**Eval-WER (%)**	**CER (%)**	**WER-Drop (%)**
ADHD only	19.48	28.84	16.05	14.28
ASD only	19.81	29.63	17.05	13.49
TD only	20.59	29.57	16.67	13.55
TD + ADHD + ASD	**16.74**	**26.43**	15.02	**16.69**
TD + ASD	21.88	26.86	**14.89**	16.26
ADHD + ASD	22.38	27.01	15.19	16.11
TD + ADHD	17.69	26.61	15.06	16.51

Note. Val-WER = Word error rate on the validation set. Eval-WER = Word error rate on the ASD test set. CER = Character error rate. WER-Drop = Relative reduction in Eval-WER from the zero-shot baseline (43.12%). All values are reported in percentages.

4.1.1 Quantitative Overview of Baseline Fine-Tuning

Among all configurations, the TD+ADHD+ASD group achieved the best performance with the lowest Word Error Rate (WER: 26.43%) and a competitive Character Error Rate (CER: 15.02%). It also demonstrated the largest relative WER reduction (16.69%) from the zero-shot baseline (43.12%).

These results confirm that including a diverse range of developmental speech profiles in the fine-tuning process improves model generalization to the atypical and heterogeneous speech patterns of autistic children.

By contrast, the TD+ASD group, despite using a larger dataset than single-group configurations, showed the lowest performance (WER: 26.86%, CER: 14.89%). While the CER was slightly lower than others, the high WER suggests that excluding ADHD speech, which may contribute prosodic diversity and disfluency variation, negatively impacts recognition performance.

Interestingly, the TD+ADHD (WER: 26.61%, CER: 15.06%) and ADHD+ASD (WER: 27.01%, CER: 15.19%) groups also achieved strong results, reinforcing the idea that ADHD speech may provide beneficial fluency variability during training.

All three single-group configurations, TD only (WER: 29.57%), ADHD only (WER: 28.84%), and ASD only (WER: 29.63%), performed comparably and worse than multi-group settings. This suggests that training on a single population is insufficient to handle the variability found in ASD speech, and further supports the advantage of heterogeneous data inclusion.

Although both WER and CER were evaluated, this study primarily focuses on Eval-WER (%) as the main performance metric. This is because WER more directly reflects errors at the lexical level, which are crucial in capturing content accuracy and communicative intent in spontaneous, atypical speech. In contrast, CER—while informative for phoneme-level fidelity—may overlook larger unit errors such as word substitutions or deletions, which often carry more serious semantic consequences.

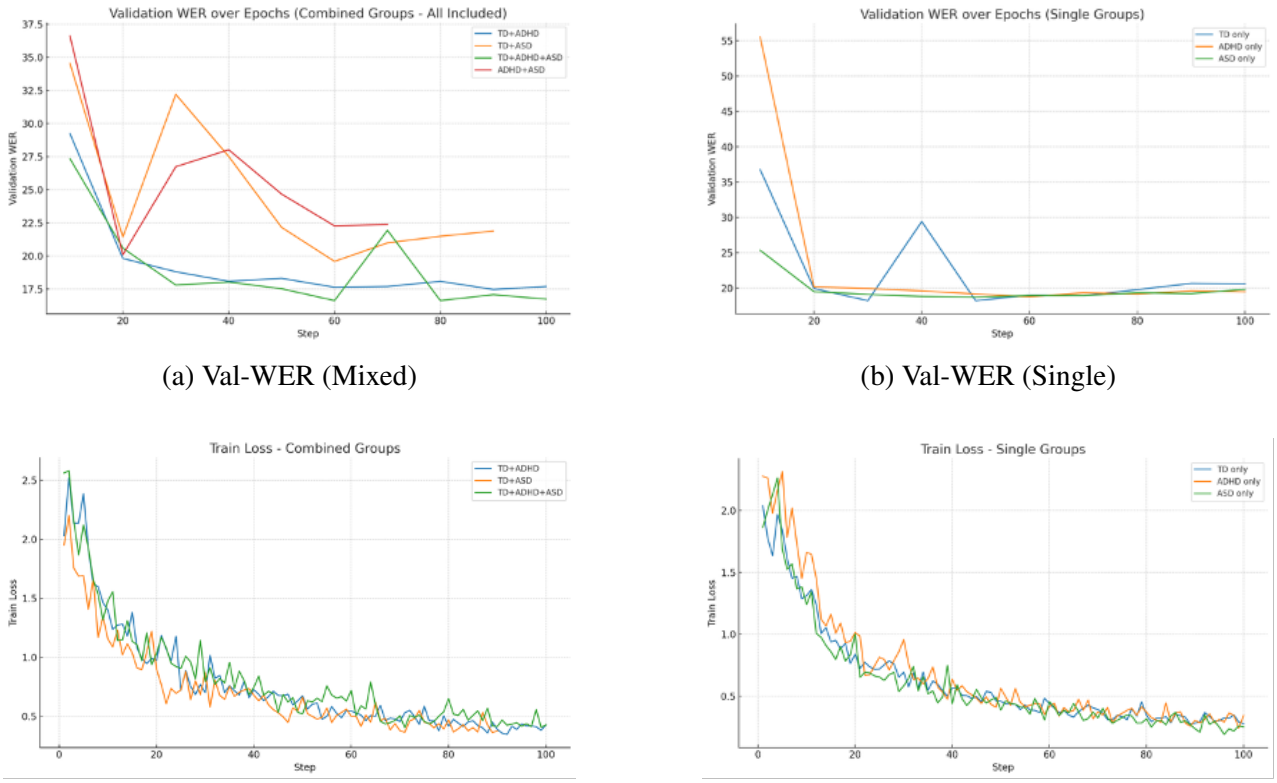
4.1.2 Training and Validation Dynamics

In addition to final evaluation scores, it is important to examine how each model learns during training. This section presents the training loss and validation performance curves across different fine-tuning configurations. Analyzing these dynamics helps assess convergence behavior, model stability, and potential overfitting, providing insight into the learning process beyond static metrics.

The training loss plots (Figure 3, c and d) show that all models converged smoothly, with most groups reaching a stable loss below 0.5 after 100 steps. While TD+ASD shows a slightly faster early decline, the final loss levels are similar across all settings.

The validation WER trends (Figure 3, a and b) highlight some important differences. The TD + ADHD + ASD configuration consistently achieves the lowest validation WER, suggesting stronger generalization. Both TD-only and ASD-only models also perform competitively, with stable and low validation WERs throughout.

In contrast, the TD+ASD configuration exhibits a more fluctuating validation WER. Despite converging quickly in training loss, this inconsistency in validation WER may reflect mismatched acoustic patterns or speaker variability between the TD and ASD groups. Prior work has shown that children with ASD often exhibit irregular prosodic timing and flatter pitch contours, which can



(a) Val-WER (Mixed)

(b) Val-WER (Single)

(c) Train Loss (Mixed)

(d) Train Loss (Single)

Figure 3: Whisper Baseline Fine-tuning Training Process

interfere with acoustic-to-linguistic mapping in ASR systems (Diehl & Paul, 2013). The ADHD-only group also shows relatively higher WER values and some instability during training, possibly due to the exaggerated pitch and intensity variability commonly found in ADHD speech (Nilsen et al., 2016), which can make temporal alignment more challenging for models trained on more stable input. However, given the general downward trend across all groups, I do not overinterpret these fluctuations.

Overall, the results suggest that the TD+ADHD+ASD configuration yields the best generalization across speech types, achieving the lowest Eval-WER and largest WER reduction. In contrast, single-group training performs worse, highlighting the limitations of narrow-domain fine-tuning. While the TD+ASD group achieves the best CER, its relatively higher WER suggests that phoneme-level accuracy does not necessarily translate into robust lexical recognition, especially in the presence of disfluencies and prosodic variation.

4.2 LoRA Fine-Tuning Results

4.2.1 Quantitative Overview of LoRA Fine-Tuning

To explore the parameter-efficient fine-tuning potential of Whisper-large, I applied LoRA (Low-Rank Adaptation) to two configurations: ASD-only and TD+ADHD+ASD due to their high per-

formance gained in the previous experiment. Table 6 presents the results in terms of Validation WER, Evaluation WER, Character Error Rate (CER), and WER reduction from the zero-shot baseline (43.12%).

Table 6: LoRA Fine-tuning Evaluation Results

Fine-tuning Group	**Val-WER (%)**	**Eval-WER (%)**	**CER (%)**	**WER-Drop (%)**
TD+ADHD+ASD(baseline fine-tuning)	16.74	26.43	15.02	16.69
TD+ADHD+ASD	20.26	38.15	24.38	4.97
ASD only(baseline fine-tuning)	19.81	29.63	17.05	13.49
ASD only	**15.5**	**23.20**	**13.34**	**19.92**

Among the two configurations of LoRA fine-tuning, the ASD-only LoRA model yielded the lower WER (23.20%) and CER (13.34%), along with the larger WER reduction (19.92%), which also performs the best among baseline fine-tuning experiments. This indicating that domain-specific tuning under LoRA constraints can be highly effective when well-aligned with the target domain.

In contrast, the TD+ADHD+ASD LoRA model resulted in higher WER (38.15%) and CER (24.38%), which is notably worse than its baseline counterpart. This performance drop may result from a mismatch between the diverse training inputs and the highly ASD-specific test set, which LoRA may not be expressive enough to accommodate.

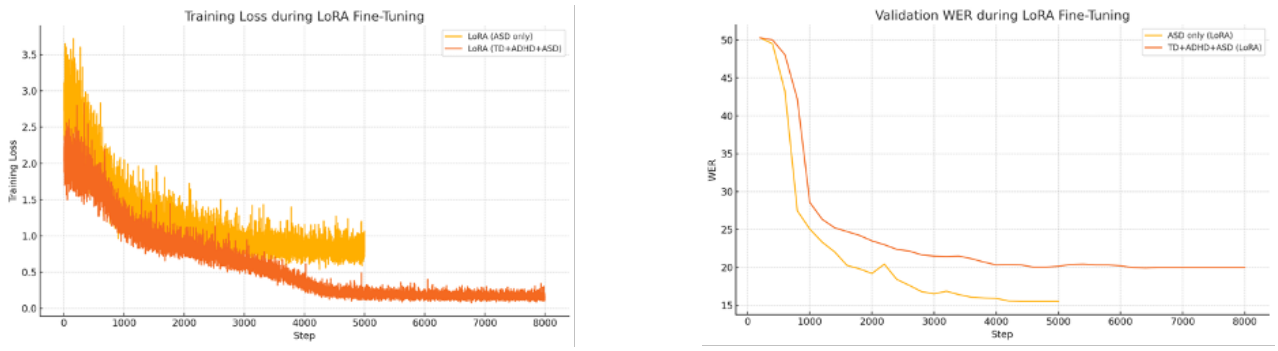
These findings suggest that LoRA performs best when the training set closely resembles the target speech profile, and its benefits diminish in more heterogeneous training contexts.

4.2.2 Training and Validation Dynamics under LoRA

Figure 4 shows that both configurations achieved consistent convergence under LoRA fine-tuning, though the TD+ADHD+ASD model converged more slowly, likely due to data complexity. The ASD-only model reached a stable low loss earlier, which aligns with its superior final performance.

Validation WER trends further reflect this distinction. The ASD-only curve rapidly drops and stabilizes below 20%, whereas the TD+ADHD+ASD curve fluctuates and plateaus around 20%, with minimal further gains after 4000 steps.

Taken together, the training dynamics suggest that while LoRA can maintain convergence, its sensitivity to domain mismatch may impact final generalization. These results illustrate that even in cases where final evaluation scores are suboptimal, LoRA still facilitates efficient convergence during training. This highlights its practical value as a lightweight fine-tuning method, especially when computational resources are limited or rapid adaptation is needed. Rather than replacing full fine-tuning in all cases, LoRA offers a promising complementary strategy—particularly effective when training data is well-aligned with the target domain.



(a) Train Loss (b) Val WER

Figure 4: Whisper LoRA Fine-tuning Training Process

4.3 Text-Based Recognition Errors Analysis

To better understand the patterns of misrecognition in ASD speech, I conducted a systematic analysis of recognition errors from four fine-tuned models and a zero-shot Whisper baseline. These five experimental conditions, TD-only, ADHD-only, ASD-only, TD+ADHD+ASD, and zero-shot, represent key contrasts in training data composition and model exposure, ranging from single-group to multi-group and from domain-matched to domain-mismatched settings. This section addresses Research Question 2 (RQ2): Are recognition errors associated with prosodic and disfluent patterns in the speech of children with ASD?

This part of analysis includes two dimensions: (1) word-level recognition errors categorized into substitutions (S), deletions (D), and insertions (I); and (2) disfluency recognition performance, specifically for filler words (e.g., ”uh”, ”um”) and word repetitions. Each analysis is based on the same 166 utterances from autistic children.

Zero-shot

Table 7: Speech Recognition Error Statistics

Error Type	**Total Count**	**Mean per Utterance**
Substitution (S)	971	5.88
Deletion (D)	282	1.71
Insertion (I)	160	0.97
Total Utterances		166

Table 8: Disfluency Recognition Accuracy

Disfluency Type	**Count in Reference**	**Count in Prediction**	**Correctly Recognized**
Filler Words	51	0	0 / 51 (0.0%)
Repetitions	30	0	0 / 30 (0.0%)

The zero-shot Whisper model, applied without any task-specific fine-tuning, performed worst among all conditions. It produced an average of 5.88 substitutions, 1.71 deletions, and 0.97 insertions per utterance. Notably, it failed to detect any filler words or repetitions, achieving 0% disfluency recognition. These results highlight the model's limited ability to generalize to ASD speech from pretraining alone, especially when prosodic irregularities are present. Prediction results are shown in appendix.

TD-only

Table 9: Speech Recognition Error Statistics

Error Type	**Total Count**	**Mean per Utterance**
Substitution (S)	692	4.17
Deletion (D)	137	0.83
Insertion (I)	142	0.86
Total Utterances		166

Table 10: Disfluency Recognition Accuracy

Disfluency Type	**Count in Reference**	**Count in Prediction**	**Correctly Recognized**
Filler Words	51	54	28 / 51 (54.9%)
Repetitions	30	22	20 / 30 (66.7%)

The TD-only model yielded slightly better performance than the zero-shot condition, averaging 4.17 substitutions, 0.83 deletions, and 0.86 insertions per utterance. While substitution errors still dominated, this model showed moderate success in disfluency recognition: 54.9% of filler words and 66.7% of repetitions were correctly identified. This suggests that exposure to typical child speech provides some transferable knowledge, particularly for recurring disfluency patterns.

ADHD-only

Table 11: Speech Recognition Error Statistics

Error Type	**Total Count**	**Mean per Utterance**
Substitution (S)	653	3.93
Deletion (D)	153	0.92
Insertion (I)	141	0.85
Total Utterances		166

Table 12: Disfluency Recognition Accuracy

Disfluency Type	Count in Reference	Count in Prediction	Correctly Recognized
Filler Words	51	53	27 / 51 (52.9%)
Repetitions	30	21	20 / 30 (66.7%)

The ADHD-trained model demonstrated a similar substitution rate (3.93) but slightly higher deletion (0.92) and comparable insertion (0.85) rates compared to the TD-only model. It detected 52.9% of filler words and 66.7% of repetitions. These results imply that ADHD speech may share some prosodic characteristics with ASD speech, enabling partial generalization, though decoding stability remains a challenge.

ASD-only

Table 13: Speech Recognition Error Statistics

Error Type	Total Count	Mean per Utterance
Substitution (S)	696	4.19
Deletion (D)	173	1.04
Insertion (I)	104	0.63
Total Utterances		166

Table 14: Disfluency Recognition Accuracy

Disfluency Type	Count in Reference	Count in Prediction	Correctly Recognized
Filler Words	51	42	24 / 51 (47.1%)
Repetitions	30	18	19 / 30 (63.3%)

Training on ASD speech directly reduced substitution errors slightly (4.19) but increased deletion rates to 1.04, the highest among all fine-tuned models. Insertion errors dropped to 0.63. This model had the lowest filler word detection rate (47.1%) but maintained a reasonably good repetition recognition rate (63.3%). These results reflect the acoustic complexity and variability within ASD speech, where even matched training data offers limited consistency.

TD-ADHD-ASD

Table 15: Speech Recognition Error Statistics

Error Type	**Total Count**	**Mean per Utterance**
Substitution (S)	575	3.46
Deletion (D)	159	0.96
Insertion (I)	134	0.81
Total Utterances		166

Table 16: Disfluency Recognition Accuracy

Disfluency Type	**Count in Reference**	**Count in Prediction**	**Correctly Recognized**
Filler Words	51	62	31 / 51 (60.8%)
Repetitions	30	22	20 / 30 (66.7%)

The multi-group model achieved the best overall balance, with the lowest substitution rate (3.46) and moderate deletion (0.96) and insertion (0.81) levels. Importantly, it showed the highest filler word detection accuracy (60.8%) and stable recognition of repetitions (66.7%). This suggests that incorporating diverse speech sources in training enhances the model's robustness to disfluent and prosodically atypical input.

Summary Comparison Across all five conditions, substitution errors were the most frequent, suggesting that lexical and phoneme-level mismatches are a persistent challenge in recognizing ASD speech.

Deletion errors were more pronounced in the zero-shot and ASD-only conditions, indicating instability in decoding atypical or unfamiliar speech patterns without sufficient exposure.

For disfluency recognition, only fine-tuned models demonstrated any ability to detect filler words or repetitions, with the multi-group (TD+ADHD+ASD) model performing best overall.

These findings suggest that training data diversity enhances recognition stability and disfluency detection. However, the current analysis is limited to recognition output patterns and does not directly examine acoustic or prosodic causes of the observed errors. Further sections incorporate acoustic-prosodic features to explore whether and how such speech characteristics may contribute to recognition difficulties.

4.3.1 Summary of Fine-Tuning Improvements Compared to Zero-Shot

Compared to the zero-shot Whisper model, the fine-tuned models achieved substantial improvements across multiple dimensions:

- Significant reduction in WER: In the baseline fine-tuning, the TD+ADHD+ASD group performs the best, reduced WER from 43.12% to 26.43% (a 16.69% relative reduction). In LoRA

fine-tuning, the ASD-only model further lowered WER to 23.20%, a 19.92% relative improvement.

- Improved CER: CER was reduced from approximately 17.8% (zero-shot) to 13.34% (ASD-only LoRA), indicating more accurate recognition at the phoneme level.
- Marked gains in disfluency recognition: The zero-shot model failed to detect any fillers or repetitions (0% accuracy).), indicating more accurate recognition at the phoneme level. The fine-tuned TD+ADHD+ASD model recognized 60.8% of fillers and 66.7% of repetitions, reflecting better retention of spontaneous speech characteristics.
- Better distribution of error types: Substitution errors decreased from 5.88 per utterance (zero-shot) to 3.46 (TD+ADHD+ASD); Deletion and insertion errors also declined notably.

Overall, the fine-tuned Whisper model outperforms the zero-shot baseline not only in error rate but also in retaining natural speech patterns. It demonstrates enhanced robustness and inclusivity, especially when dealing with spontaneous, neurodivergent child speech.

4.4 Prosodic and Acoustic-Based Recognition Errors

While text-based error analysis reveals surface-level mismatches between ground truth and ASR outputs, it offers limited insight into the acoustic or prosodic patterns that may underlie recognition difficulties. To further investigate whether such speech-level characteristics are associated with recognition accuracy, a targeted analysis was conducted on utterances exhibiting extreme recognition outcomes. Specifically, for each of the four fine-tuned models (TD-only, ADHD-only, ASD-only, and TD+ADHD+ASD), the three most accurately and the three least accurately transcribed utterances were identified, based on WER and CER.

For each of these 24 utterances, I extracted a set of prosodic and voice-quality features, including:

- Pitch Standard Deviation (Hz)
- Pitch Range (Hz)
- Voiced Ratio (
- Speech Rate (frames/second)
- Mean Intensity (dB)
- Intensity Standard Deviation (dB)
- Shimmer (local)

These features were selected to reflect core dimensions of speech prosody, vocal quality, and articulation dynamics. While pitch- and rate-based measures capture prosodic variability, shimmer and intensity variation offer additional information about phonatory stability and loudness modulation.

By comparing the features of the most and least accurately recognized utterances, I aimed to identify

patterns potentially associated with recognition performance.

TD-only

The three most error-prone utterances were asd10_007, asd38_010, and asd38_012, while the most accurately transcribed samples were asd02_004, asd22_010, and asd46_010. Key contrasts included a much higher pitch variation in the accurate group (Pitch SD = 61.56 Hz vs. 12.28 Hz; Pitch Range = 363.39 Hz vs. 55.13 Hz), greater intensity variability (Intensity SD = 5.73 dB vs. 3.14 dB), and lower shimmer values, suggesting that monotonous and acoustically flat utterances tend to yield more recognition errors.

Table 17: Prosodic Feature Comparison

Feature	**Top 3 most accurate**	**Top 3 least accurate**
Pitch SD (Hz)	61.56	12.28
Pitch Range (Hz)	363.39	55.13
Voiced Ratio	43.6%	44.2%
Speech Rate (voiced/sec)	86.89	87.45
Intensity Mean (dB)	74.61	73.05
Intensity SD (dB)	5.73	3.14
Shimmer (local)	0.154	0.176

ADHD-only

Among the ADHD-trained model, the least accurate samples were asd38_011, asd38_012, and asd12_002, while the best were asd02_008, asd22_004, and asd43_002. Accurate utterances had higher pitch variation (Pitch SD = 55.77 Hz vs. 28.53 Hz), slightly slower speech rate, and marginally greater mean intensity. Notably, shimmer values were nearly identical, indicating that speech tempo and pitch modulation may matter more than voice quality in this condition. **ADHD-only**

Table 18: Prosodic Feature Comparison

Feature	**Top 3 Most Accurate**	**Top 3 Least Accurate**
Pitch SD (Hz)	55.77	28.53
Pitch Range (Hz)	218.27	173.42
Voiced Ratio	43.7%	49.8%
Speech Rate (frames/s)	86.82	97.58
Mean Intensity (dB)	75.27	74.07
Intensity SD (dB)	5.25	5.36
Shimmer (local)	0.152	0.151

ASD-only

For the ASD-only model, high-error utterances included asd19_004, asd12_002, and asd38_010, and the most accurate were asd02_008, asd02_009, and asd19_009. Here, the most reliable differentiator was voiced ratio (50.5% vs. 25.4%), alongside significantly faster speech rate and higher mean intensity in the accurate group. This suggests that utterances with sparse voicing and slow articulation may be particularly challenging to recognize, even for models trained on ASD speech. **ASD-only**

Table 19: Prosodic Feature Comparison

Feature	Top 3 Most Accurate	Top 3 Least Accurate
Pitch SD (Hz)	59.46	60.49
Pitch Range (Hz)	250.04	359.94
Voiced Ratio	50.5%	25.4%
Speech Rate (frames/s)	99.92	50.33
Mean Intensity (dB)	75.42	72.53
Intensity SD (dB)	5.08	4.93
Shimmer (local)	0.144	0.142

TD+ADHD+ASD

The combined training model's lowest-performing samples were asd10_007, asd12_002, and asd38_010, while its highest-performing ones were asd02_004, asd22_003, and asd19_001. Accurate utterances showed higher pitch variability (Pitch SD = 57.12 Hz vs. 41.19 Hz), more voiced frames, and greater intensity variation. These patterns again point to the importance of dynamic prosody and consistent voicing in facilitating recognition. **TD+ADHD+ASD**

Table 20: Prosodic Feature Comparison

Feature	Top 3 Most Accurate	Top 3 Least Accurate
Pitch SD (Hz)	57.12	41.19
Pitch Range (Hz)	267.32	269.45
Voiced Ratio	37.0%	32.3%
Speech Rate (frames/s)	73.67	64.14
Mean Intensity (dB)	73.59	74.85
Intensity SD (dB)	5.89	4.46
Shimmer (local)	0.157	0.142

Across all four models, certain recurring trends emerged: utterances with greater pitch variability, higher intensity dynamics, and more consistent voicing were more likely to be transcribed accurately. Conversely, samples with low pitch variation, reduced voicing, or flatter intensity profiles appeared among the least accurately recognized. However, the small sample size and descriptive nature of this analysis limit the generalizability of these findings. While the observed feature contrasts are suggestive, they do not establish causal relationships between prosodic features and ASR performance. To address this limitation, the next section presents a large-scale, quantitative analysis that aggregates feature values across all ASD test utterances, enabling statistical inference about their associations with recognition error rates.

4.4.1 Aggregated Analysis of Prosodic Features and Recognition Errors

Building upon the previous case-level comparisons, this section offers a broader, statistical perspective by aggregating prosodic features across all ASD utterances evaluated by the TD+ADHD+ASD model, the most accurate system identified in prior experiments. The objective is to determine whether specific prosodic and acoustic-prosodic features systematically correlate with recognition outcomes, particularly with respect to word error rate (WER) and character error rate (CER).

Correlation Analysis

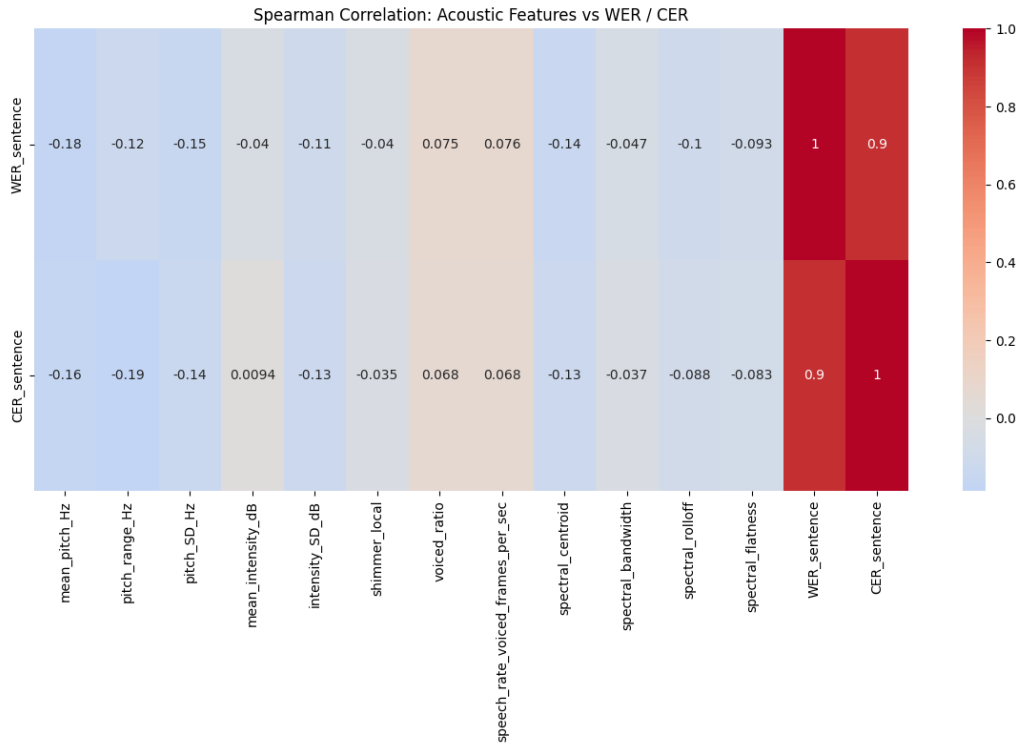


Figure 5: Spearman Heatmap

To examine monotonic relationships between prosodic features and recognition errors, Spearman's rank correlation coefficients (ρ) were computed between 12 acoustic-prosodic features and sentence-level WER and CER scores. The results, visualized in a Spearman correlation heatmap (Figure), revealed mostly weak to moderate negative correlations for pitch-related features. The results, visualized in a Spearman correlation heatmap (Figure X), revealed mostly weak to moderate negative correlations for pitch-related features.

Spearman's rank correlation coefficient (ρ) is a statistical method that measures whether two variables increase or decrease together in a consistent (monotonic) way. It does not assume a linear relationship and is useful for analyzing speech data, which can be noisy or unevenly distributed. The prosodic features include measures like pitch variation, intensity, shimmer (voice stability), and voiced ratio, which are known to affect how clearly speech is recognized by ASR systems.

Pitch-related features demonstrated the most consistent inverse relationships:

- Pitch range (ρ = -0.19 with CER, -0.12 with WER)
- Pitch standard deviation (ρ = -0.14 with CER, -0.15 with WER)

These results suggest that flatter intonation, characterized by reduced pitch variability, is modestly associated with increased recognition errors, reinforcing the observations from case-level analysis. In contrast, voiced ratio and speech rate showed weak positive correlations with WER and CER ($\rho \approx 0.07$), possibly reflecting higher error risk in rapid or densely voiced speech. Other features, such as shimmer and spectral flatness, exhibited minimal correlation ($|\rho| < 0.1$), suggesting a limited direct impact on recognition performance.

Multiple Linear Regression

To further explore the combined effects of prosodic features, an ordinary least squares (OLS) regression was performed with sentence-level WER as the dependent variable and the 12 prosodic-acoustic features as predictors.

The regression results (Figure) identified voiced ratio as a significant positive predictor of WER (β = 2183.18, p = 0.014), suggesting that a higher proportion of voiced frames may complicate recognition. Conversely, speech rate had a significant negative coefficient (β = -11.03, p = 0.014), indicating that slower articulation is associated with higher recognition error.

While pitch-related features such as shimmer and spectral flatness had large coefficients, they did not reach statistical significance, possibly due to multicollinearity or limited explanatory power. Multicollinearity refers to high intercorrelation between predictors, which can inflate coefficient values and reduce the reliability of significance testing, making it difficult to isolate the independent contribution of each feature. Limited explanatory power, on the other hand, refers to the relatively low

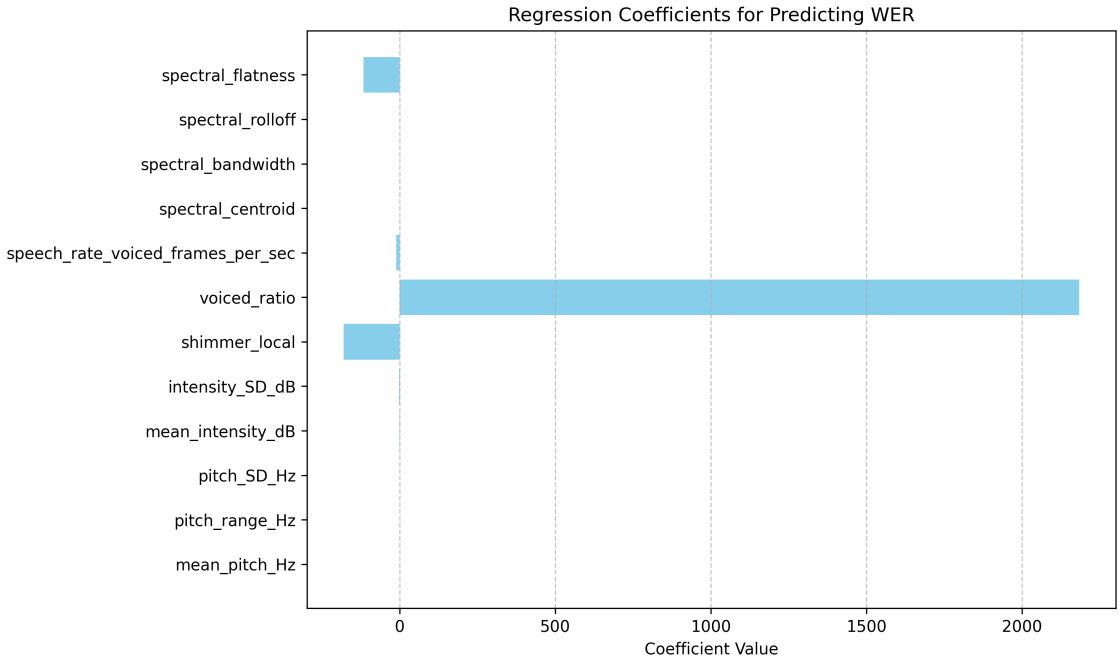


Figure 6: Multiple Regression

proportion of variance in WER explained by the model.

The model accounted for approximately 13% of the variance in WER ($R^2 = 0.13$), indicating that while prosodic features influence recognition, other factors, such as linguistic complexity, phonetic content, or background noise, also contribute.

Group-Wise Comparison: High-WER vs Low-WER

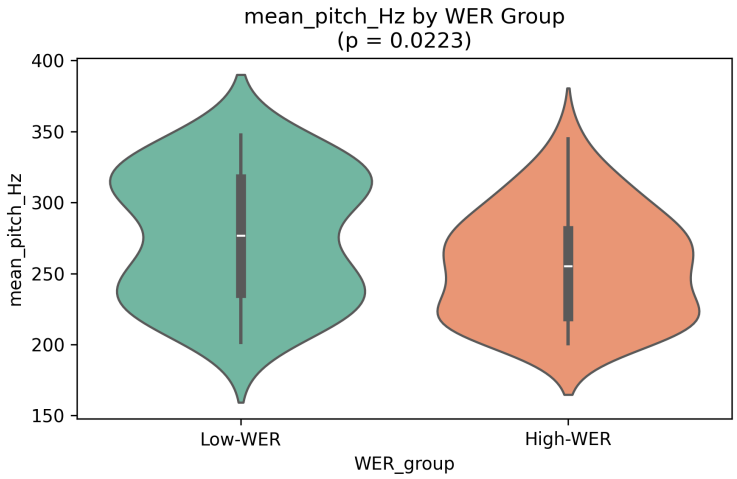


Figure 7: Mean Pitch Hz Violin

Table 21: Mann–Whitney U Test Results for Acoustic Features

Features	**Mann–Whitney U p-value**
mean_pitch_Hz	**0.022**
pitch_range_Hz	0.153
pitch_SD_Hz	0.161
mean_intensity_dB	0.196
intensity_SD_dB	0.345
shimmer_local	0.671
voiced_ratio	0.782
speech_rate_voiced_frames_per_sec	0.899
spectral_centroid	0.782
spectral_bandwidth	0.671
spectral_rolloff	0.614
spectral_flatness	0.486

To statistically validate specific feature differences, ASD utterances were divided into two subsets based on WER distribution: top 25% (high-WER) and bottom 25% (low-WER). A Mann-Whitney U test was conducted to assess whether any of the 12 prosodic features significantly differed between the two groups.

Among all tested features, only mean pitch exhibited a statistically significant difference ($p = 0.022$). As shown in Figure 7, low-WER utterances had higher mean pitch values than high-WER utterances. This result aligns with earlier findings and supports the interpretation that flatter intonation, reflected by lower average pitch, may systematically hinder ASR performance on ASD speech.

Together, the three levels of analysis, correlation, regression, and group-wise comparison, offer converging evidence that pitch-related features, particularly pitch variation and mean pitch, are closely associated with ASR of ASD children speech. Consistent with prior literature on prosodic flattening in ASD, the results suggest that reduced pitch dynamics may present systematic challenges for automatic recognition systems.

Meanwhile, features such as speech rate and voiced ratio demonstrated model-dependent effects, reflecting interactions with temporal segmentation and acoustic encoding. In sum, while prosodic features alone cannot fully explain recognition performance, they represent a critical layer that future ASR systems have to address to include neurodiverse speakers.

5 Discussion and Conclusion

5.1 Summary of Key Findings

This study set out to address two main research questions:

(1) whether fine-tuning Whisper improves ASR performance on Dutch speech from children with ASD, and

(2) whether prosodic and disfluency features correlate with recognition errors.

In response to RQ1, whether fine-tuning improves ASR performance for Dutch-speaking autistic children, the results provide clear evidence that both baseline and parameter-efficient fine-tuning substantially reduce word error rates (WER) compared to zero-shot performance. Among baseline fine-tuning configurations, the TD+ADHD+ASD group yielded the strongest generalization performance, achieving a WER of 26.43%, indicating that exposure to diverse developmental speech patterns enhances robustness. Complementarily, the LoRA fine-tuning experiments revealed that even with ASD-only data, the model achieved the lowest WER overall (23.20%), surpassing all baseline results. This suggests that small but domain-aligned datasets, when combined with efficient adaptation techniques, can yield significant gains in ASR performance for neurodivergent speech. Together, these findings demonstrate that targeted fine-tuning can meaningfully improve recognition accuracy for autistic children's speech, even under low-resource constraints.

Regarding RQ2, the study conducted both qualitative and quantitative error analyses to further explore the relationship between recognition errors and prosodic/disfluency features. These included word-level error categorization (substitutions, deletions, insertions), disfluency recognition accuracy (e.g., fillers, repetitions), and statistical analyses (e.g., Mann-Whitney U tests, regression modeling). The analyses revealed that key prosodic features, such as mean pitch, pitch variability, speech rate, and voicing proportion, were significantly correlated with WER outcomes. From these results, I concluded that prosodic irregularities and disfluencies play a central role in driving ASR errors, particularly in zero-shot settings. Fine-tuned models, especially those trained on multi-group data, preserved more disfluencies and reduced deletion/substitution rates. However, some outliers were observed: for example, the ASD-only model exhibited higher deletion errors despite domain alignment, which may reflect intra-group variability or speaker-level idiosyncrasies. This aligns with previous literature emphasizing the acoustic heterogeneity within the ASD population, suggesting that even domain-specific models may benefit from more diverse or individualized training strategies.

In the following subsections, the results will be contextualized through comparison with prior literature, critical discussion of methodological limitations, and proposals for future research directions. By situating the findings within the broader landscape of speech technology and clinical ASR research, this section aims to clarify the implications and potential impact of the current study.

5.2 Comparison with Previous Research

This study's findings both confirm and extend several strands of prior research outlined in the literature review.

First, the observed improvement in WER and CER after fine-tuning supports earlier work suggesting that ASR systems can benefit from domain adaptation when dealing with non-typical speech. For example, Tobin et al. (2024) and Lee et al. (2024) highlighted the limitations of off-the-shelf models on neurodivergent children's speech and advocated for task-specific training. Our results reinforce this, showing that even lightweight fine-tuning methods like LoRA lead to substantial WER reductions (from 43.12% to 23.20%), thereby validating prior claims.

Second, this study builds on and adds empirical depth to the work of Asghari et al. (2021); Liu et al. (2024), who hypothesized that prosodic irregularities in ASD speech, such as flattened intonation and altered speech rate, are a key cause of recognition failures. By statistically correlating prosodic features (e.g., pitch variance, speech rate, voicing ratio) with recognition errors, our study provides concrete evidence for these assumptions, which were previously more theoretical or qualitative. Third, unlike Gale et al. (2019) who mainly relied on read speech and observed moderate gains from fine-tuning, this study uses spontaneous speech from Dutch-speaking ASD children and still achieves significant error reduction. This suggests that spontaneous speech may contain more prosodic cues useful for model adaptation, especially when combined with multi-group training data (TD+ADHD+ASD).

However, one finding that partially diverges from prior work is the unexpectedly strong performance of ASD-only fine-tuning under LoRA, which contrasts with literature suggesting broader data diversity leads to better generalization (e.g., Q. Liu et al. (2024)). A explanation is that, in our low-resource setting, domain-specific consistency (ASD-to-ASD) outweighs generalization benefits, especially when test and training data are matched.

In summary, this study largely supports existing findings while providing quantitative verification and new nuance. It emphasizes that successful ASR adaptation for ASD speech must consider both acoustic-phonetic irregularities and data alignment strategies.

5.3 Limitations and future research

While this study offers valuable insights into ASR performance on Dutch speech from children with ASD, several limitations should be acknowledged across the dataset, model design, and experimental setup.

Corpus Limitations

First, the ASDBank corpus used in this study is limited in size, comprising approximately 2.4 hours of speech data in total, with restricted speaker diversity and content variation, averaging only about two minutes of speech per speaker. Although this study applied speaker-independent splits, the small number of speakers constrains the generalizability of findings. In addition, the dataset includes only

children with mild to moderate ASD symptoms and consists largely of picture description tasks, which may not capture the full diversity of speech styles or disfluency patterns in natural conversation.

Prosodic Feature Extraction and Analysis

Second, the analysis of prosodic features was conducted post hoc, using Praat to extract averaged values for each audio segment. This approach may oversimplify temporal dynamics or interactions between features (e.g., pitch and speech rate). Furthermore, while this study focused on a subset of prosodic features, additional descriptors such as jitter, shimmer, and pause duration were not included. Future work could benefit from incorporating these features, as they may capture relevant acoustic cues associated with recognition errors and further enhance model performance in neurodivergent speech contexts.

Additionally, not all statistical comparisons employed standard parametric tests such as the t-test. For instance, Mann-Whitney U tests were used to compare prosodic patterns between high- and low-WER utterance groups, due to the non-normal distribution and small sample size of acoustic features. Spearman's rank correlation and OLS regression were used to explore associations between prosodic features and recognition error rates. However, no t-tests were conducted to directly compare performance metrics (e.g., WER) across all training conditions or experimental groups. Future studies with larger and more balanced datasets could incorporate parametric tests to enable more robust statistical comparisons and confidence estimates. **Model Configuration and Evaluation**

Although the use of LoRA fine-tuning helped improve performance under low-resource settings, the model was only partially adapted by modifying parameters in the higher layers. This narrow parameter space might have limited the model's ability to adapt to complex ASD-specific prosodic patterns. Additionally, the training steps (70-100) and small batch size (16) may have underutilized the learning capacity of the Whisper-medium model.

In terms of evaluation, the primary metric used was Word Error Rate (WER), without incorporating phoneme-level analysis or listener-based intelligibility assessments. Expanding the evaluation framework to include these additional dimensions in future work could help capture more subtle improvements in speech naturalness and semantic preservation that may not be reflected by WER alone.

Explanation for Suboptimal Results in Some Settings

Finally, models fine-tuned solely on TD or ADHD data underperformed, likely due to differences in prosodic profiles and disfluency patterns relative to the ASD test set. Similarly, the LoRA fine-tuned model trained on the combined TD+ADHD+ASD data also showed suboptimal results, due to its limited adaptation capacity in handling heterogeneous input profiles. These findings underscore the importance of both domain-relevant data and training-target alignment, especially when employing parameter-efficient methods like LoRA for neurodivergent speech modeling.

5.4 Future Work

Based on the limitations discussed above, several directions for future research are proposed, spanning dataset expansion, model improvements, and advanced evaluation methods, many of which are grounded in findings from prior literature.

Expanding Dataset Diversity and Task Complexity

To address the current corpus limitation, future work should collect larger, more diverse datasets of Dutch children with ASD. This includes:

- Speech samples from children with a broader range of ASD severity, including those with co-occurring language impairments.
- A variety of speaking contexts (e.g., free conversation, storytelling, interactional dialogue) beyond picture descriptions, to capture a richer array of disfluency types and prosodic dynamics.
- Inclusion of more speakers to improve generalization and reduce overfitting risks.

These expansions would help model more realistic and complex communication behaviors, as advocated by Patel et al. (2023) and Plate (2025).

Improving Prosody-Aware Modeling

As suggested in the literature (e.g., Sohn, Knutsen, and Stromswold (2025)), ASR performance for neurodivergent speakers can benefit from explicitly modeling prosodic and acoustic markers. Future models could incorporate:

- Multimodal input that fuses raw speech with extracted prosodic features (pitch, energy, pause, speech rate).
- Feature-conditioned decoding, where the ASR decoder attends to rhythmic and prosodic contours.
- Integration with emotion or stress detection models to better capture atypical speech modulations in ASD.

Furthermore, freezing fewer layers in the model during fine-tuning (i.e., beyond q_proj and v_proj) and using more targeted parameter-efficient tuning (e.g., LoRA+PEFT) may enable better adaptation to the ASD-specific prosodic patterns.

Improved Error and Disfluency Analysis through Visualization

A particularly promising direction is the visualization of word-level recognition aligned with prosodic timelines. This could include:

- Dynamic word-pitch overlays (e.g., pitch contours aligned with transcript timelines).

- Disfluency heatmaps showing where fillers, hesitations, and self-corrections occur relative to recognition errors.
- Comparison dashboards of Whisper predictions vs ground truth with aligned acoustic-prosodic patterns.

Such visualization tools would not only enhance interpretability of model errors but also support clinical applications, such as helping therapists or educators analyze speech in therapeutic settings.

5.5 Main Contributions

Collectively, the findings presented in this thesis provide experimental evidence that fine-tuning not only reduces surface-level transcription errors but also enhances model sensitivity to the nuanced prosodic and disfluency characteristics of speech produced by children with Autism Spectrum Disorder. These improvements were observed even under low-resource constraints, using small, domain-relevant datasets.

Beyond the technical contributions, this study offers meaningful implications for both academic research and practical applications. From an academic perspective, it extends the current understanding of automatic speech recognition performance on neuron-degenerative child speech, particularly in underrepresented languages such as Dutch. By integrating LoRA-based fine-tuning with detailed prosodic analysis, the study demonstrates the feasibility of adapting large-scale ASR models like Whisper to atypical speech domains, establishing a replicable and scalable pipeline for future research involving disordered or diverse speech populations.

From a practical standpoint, the improved recognition accuracy on Dutch ASD speech lays the foundation for more inclusive and accessible speech technologies. Potential applications include digital therapeutic tools, personalized educational platforms, and speech feedback systems that can better support the needs of neurodivergent children. Furthermore, this study highlights the critical role of domain-specific training data and interpretability tools—such as prosody-informed visualizations, in facilitating real-world deployment within healthcare and education contexts.

References

Asghari, et al. (2021, November). Distinctive prosodic features of people with autism spectrum disorder: A systematic review and meta-analysis study. *Scientific Reports*, *11*(1), 23093. doi: 10.1038/s41598-021-02487-6

Ashvin, A., Lahiri, R., Kommineni, A., Bishop, S., Lord, C., Kadiri, S. R., & Narayanan, S. (2024, September). *Evaluation of state-of-the-art ASR Models in Child-Adult Interactions* (No. arXiv:2409.16135). arXiv. doi: 10.48550/arXiv.2409.16135

Bhardwaj, et al. (2021, June). Usage of Prosody Modification and Acoustic Adaptation for Robust Automatic Speech Recognition (ASR) System. *Revue d'Intelligence Artificielle*, *35*(3), 235–242. doi: 10.18280/ria.350307

Bhardwaj, et al. (2022, January). Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review. *Applied Sciences*, *12*(9), 4419. doi: 10.3390/app12094419

C, et al. (2018, August). Autism spectrum disorder. *Lancet (London, England)*, *392*(10146). doi: 10.1016/S0140-6736(18)31129-2

Cassol-Jr, O. J., Comim, C. M., Silva, B. R., Hermani, F. V., Constantino, L. S., Felisberto, F., … Dal-Pizzol, F. (2010, August). Treatment with cannabidiol reverses oxidative stress parameters, cognitive impairment and mortality in rats submitted to sepsis by cecal ligation and puncture. *Brain Research*, *1348*, 128–138. doi: 10.1016/j.brainres.2010.06.023

Diehl, J. J., & Paul, R. (2013, January). Acoustic and perceptual measurements of prosody production on the profiling elements of prosodic systems in children by children with autism spectrum disorders. *Applied Psycholinguistics*, *34*(1), 135–161. doi: 10.1017/S0142716411000646

Diehl, J. J., Watson, D., Bennetto, L., Mcdonough, J., Gunlogson, et al. (2009, July). An acoustic analysis of prosody in high-functioning autism. *Applied Psycholinguistics*, *30*(3), 385–404. doi: 10.1017/S0142716409090201

Filipe, et al. (2014, August). Atypical Prosody in Asperger Syndrome: Perceptual and Acoustic Measurements. *Journal of Autism and Developmental Disorders*, *44*(8), 1972–1981. doi: 10.1007/s10803-014-2073-2

Fuckner, et al. (2023, October). Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch speakers. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (pp. 146–151). Bucharest, Romania: IEEE. doi: 10.1109/SpeD59241.2023.10314895

Gale, R., Chen, L., Dolata, J., van Santen, J., Asgari, et al. (2019, September). Improving ASR Systems for Children with Autism and Language Impairment Using Domain-Focused DNN Transfer Techniques. *Interspeech*, *2019*, 11–15. doi: 10.21437/Interspeech.2019-3161

Ghimire, et al. (2024, December). Improving on the Limitations of the ASR Model in Low-Resourced Environments Using Parameter-Efficient Fine-Tuning. In S. Lalitha Devi & K. Arora (Eds.), *Proceedings of the 21st International Conference on Natural Language Processing (ICON)* (pp. 408–415). AU-KBC Research Centre, Chennai, India: NLP Association of India (NLPAI).

Godin, K. W., & Hansen, J. H. L. (2015, October). Physical task stress and speaker variability in voice quality. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*(1), 29. doi: 10.1186/s13636-015-0072-7

Goldwater, et al. (2010, March). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, *52*(3),

181–200. doi: 10.1016/j.specom.2009.10.001

Graham, C., & Roll, N. (2024, February). Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA express letters*, *4*(2), 025206. doi: 10.1121/10.0024876

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, *23*(8), 639–652. doi: 10.1016/j.tics.2019.05.006

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021, June). *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units* (No. arXiv:2106.07447). arXiv. doi: 10.48550/arXiv.2106.07447

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021, October). *LoRA: Low-Rank Adaptation of Large Language Models* (No. arXiv:2106.09685). arXiv. doi: 10.48550/arXiv.2106.09685

Hu, Y., Zhang, C., Shi, J., Lian, J., Ostendorf, M., & Yu, D. (2022). Prosodybert: Self-supervised prosody representation for style-controllable tts.

Hubbard, et al. (2017). Production and perception of emotional prosody by adults with autism spectrum disorder. *Autism Research*, *10*(12), 1991–2001. doi: 10.1002/aur.1847

Jain, R. o. (2023, July). *Adaptation of Whisper models to child speech recognition* (No. arXiv:2307.13008). arXiv. doi: 10.48550/arXiv.2307.13008

Kohler, K. J. (2017). Speech communication in human interaction. In *Communicative functions and linguistic forms in speech interaction* (p. 1–17). Cambridge University Press.

Kuijper, et al. (2015, July). Who Is He? Children with ASD and ADHD Take the Listener into Account in Their Production of Ambiguous Pronouns. *PLOS ONE*, *10*(7), e0132408. doi: 10.1371/journal.pone.0132408

Lake, et al. (2011, February). Listener vs. speaker-oriented aspects of speech: Studying the disfluencies of individuals with autism spectrum disorders. *Psychonomic Bulletin & Review*, *18*(1), 135–140. doi: 10.3758/s13423-010-0037-x

Lee, S., Mun, J., Kim, S., Park, H., Yang, S., Kim, H., ... Chung, M. (2024). Automatic Speech Recognition and Assessment Systems Incorporated into Digital Therapeutics for Children with Autism Spectrum Disorder. In K. Miesenberger, P. Peňáz, & M. Kobayashi (Eds.), *Computers Helping People with Special Needs* (pp. 328–335). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-62849-8_40

Lehnert-LeHouillier, et al. (2020, October). Prosodic Entrainment in Conversations of Verbal Children and Teens on the Autism Spectrum. *Frontiers in Psychology*, *11*. doi: 10.3389/fpsyg.2020.582221

Liu, et al. (2024, June). Exploration of Whisper fine-tuning strategies for low-resource ASR. *EURASIP Journal on Audio, Speech, and Music Processing*, *2024*(1), 29. doi: 10.1186/s13636-024-00349-3

Liu, Q., Wu, X., Zhao, X., Zhu, Y., Xu, D., Tian, F., & Zheng, Y. (2024). When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th international acm sigir conference on research and development in information retrieval* (pp. 1104–1114).

Malik, M., et al. (2021, March). Automatic speech recognition: A survey. *Multimedia Tools and Applications*, *80*(6), 9411–9457. doi: 10.1007/s11042-020-10073-7

Mujtaba, et al. (2024, May). *Lost in Transcription: Identifying and Quantifying the Accuracy Biases of Automatic Speech Recognition Systems Against Disfluent Speech* (No. arXiv:2405.06150).

arXiv. doi: 10.48550/arXiv.2405.06150

Mujtaba, D., Mahapatra, N. R., Arney, M., Yaruss, J. S., Herring, C., & Bin, J. (2024, September). Inclusive ASR for Disfluent Speech: Cascaded Large-Scale Self-Supervised Learning with Targeted Fine-Tuning and Data Augmentation. In *Interspeech 2024* (pp. 1275–1279). doi: 10.21437/Interspeech.2024-2246

Müller-Eberstein, et al. (2024, September). Hypernetworks for Personalizing ASR to Atypical Speech. *Transactions of the Association for Computational Linguistics*, *12*, 1182–1196. doi: 10.1162/tacl_a_00696

Nilsen, et al. (2016, August). Mother-Child Communication: The Influence of ADHD Symptomatology and Executive Functioning on Paralinguistic Style. *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.01203

Olivati, et al. (2017, April). Análise acústica do padrão entoacional da fala de indivíduos com Transtorno do Espectro Autista. *CoDAS*, *29*, e20160081. doi: 10.1590/2317-1782/20172016081

Patel, S., Landau, E., Martin, G., Rayburn, C., Elahi, S., Fragnito, G., & Losh, M. (2023, 02). A profile of prosodic speech differences in individuals with autism spectrum disorder and first-degree relatives. *Journal of communication disorders*, *102*, 106313. doi: 10.1016/j.jcomdis.2023.106313

Paul, R., Bianchi, N., Augustyn, A., Klin, A., Volkmar, F. R., et al. (2008, January). Production of syllable stress in speakers with autism spectrum disorders. *Research in Autism Spectrum Disorders*, *2*(1), 110–124. doi: 10.1016/j.rasd.2007.04.001

Plate, S. N. (2025, May). The State of Natural Language Sampling in Autism Research: A Scoping Review. *Autism & Developmental Language Impairments*, *10*, 23969415251341247. doi: 10.1177/23969415251341247

Prabhavalkar, R., Hori, T., Sainath, T. N., Schlüter, R., Watanabe, et al. (2024). End-to-End Speech Recognition: A Survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *32*, 325–351. doi: 10.1109/TASLP.2023.3328283

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022, December). *Robust Speech Recognition via Large-Scale Weak Supervision* (No. arXiv:2212.04356). arXiv. doi: 10.48550/arXiv.2212.04356

Rijal, et al. (2024, November). *Whisper Finetuning on Nepali Language* (No. arXiv:2411.12587). arXiv. doi: 10.48550/arXiv.2411.12587

Santen, V., et al. (2010, May). Computational prosodic markers for autism. *Autism: The International Journal of Research and Practice*, *14*(3), 215–236. doi: 10.1177/1362361309363281

Sobti, et al. (2024, March). Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges. *Multimedia Tools and Applications*, *83*(35), 81933–81995. doi: 10.1007/s11042-024-18753-4

Sohn, S. S., Knutsen, S., & Stromswold, K. (2025, March). *Fine-Tuning Whisper for Inclusive Prosodic Stress Analysis* (No. arXiv:2503.02907). arXiv. doi: 10.48550/arXiv.2503.02907

Soleymanpour, et al. (2022, May). Synthesizing Dysarthric Speech Using Multi-Speaker Tts For Dysarthric Speech Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7382–7386). doi: 10.1109/ICASSP43922.2022.9746585

Stoyanchev, et al. (2012, December). Localized detection of speech recognition errors. In *2012 IEEE Spoken Language Technology Workshop (SLT)* (pp. 25–30). doi: 10.1109/SLT.2012.6424164

Tobin, J., Nelson, P., MacDonald, B., Heywood, R., Cave, R., Seaver, K., ... Green, J. R. (2024, November). Automatic Speech Recognition of Conversational Speech in Individuals With Disordered Speech. *Journal of Speech, Language, and Hearing Research*, *67*(11), 4176–4185. doi: 10.1044/2024_JSLHR-24-00045

Vicsi, K., & Szaszák, G. (2010, May). Using prosody to improve automatic speech recognition. *Speech Communication*, *52*(5), 413–426. doi: 10.1016/j.specom.2010.01.003

Yang, H., Zhang, M., Tao, S., Ma, M., & Qin, Y. (2023). Chinese asr and ner improvement based on whisper fine-tuning. In *2023 25th international conference on advanced communication technology (icact)* (pp. 213–217).

Zhang, D., Feng, T., Xue, L., Wang, Y., Dong, Y., & Tang, J. (2025, January). *Parameter-Efficient Fine-Tuning for Foundation Models* (No. arXiv:2501.13787). arXiv. doi: 10.48550/arXiv.2501.13787

Zorić, V. (2024, February). Pragmatic function of speech disfluencies in high-functioning children with autism spectrum disorder. *Govor/Speech*, *40*(2), 169–192. doi: 10.22210/govor.2023.40.10

Appendices

A Declaration of AI Use

During the preparation of this thesis, I used *ChatGPT (OpenAI, GPT-4, 2025)* to support the development and presentation of this work in the following ways:

- Improving academic writing quality by refining grammar, clarity, and formal tone across all chapters.
- Assisting with Python implementation for parameter-efficient fine-tuning of Whisper (LoRA), including restructuring scripts to reduce memory usage and improve reproducibility.
- Offering guidance on the interpretation of ASR metrics such as WER and CER, and how to format them consistently across different model configurations.
- Providing support in structuring prosodic error analysis results, including identifying meaningful patterns and organizing appendix tables.
- Helping format and generate LaTeX-compatible tables for prosodic summaries and prediction comparisons across conditions.

All AI-generated suggestions were critically reviewed and revised by me. The design of the experiments, interpretation of the results, and final conclusions reflect my own work and judgment. I take full responsibility for the content presented in this thesis.

Name: Hantao Yu

Date: 11.06.2025

B Prediction Results

TD only

Utterance	Reference	Prediction	WER_sentence	CER_sentence
asd10_007	piraat wil gaan voetballen	de bhas wil kan komt van	125	57.69
asd38_010	ik weet niet wat te zeggen wat zeggen	ik ga me nu op de zegger op de zegger	112.5	51.35
asd38_012	en een kraan	nokraat	100	58.33
asd02_004	de zuster plukt een bloem de zuster geeft de bloem aan de danseres zet de bloem in zn haar	de zuster plukt een bloem de zuster geeft de bloem aan de danseres de danseres zet de bloem in zn haar	5	3.03
asd22_010	de ridder heeft een netje de piraat kijkt naar t netje de ridder vist de bal uit het water de piraat wordt blij de piraat is blij met zn bal	de ridder heeft een netje de piraat kijkt naar de netje de ridder vist de bal uit het water de piraat wordt blij de piraat is blij met zn bal	3.33	1.43
asd46_010	ze doet de water in de gieter ze gaat de bloemetjes water geven en daar komt de zuster aan en die plukt een bloem en ze geeft de bloem aan de ballerina de ballerina doet t in haar haar	ze doet de water in de gieter ze gaat de bloemetjes water geven en daar komt de zuster aan en die plukt een bloem en ze geeft de bloem aan de ballerina de ballerina doet het in haar haar	2.56	1.09

ADHD only

Utterance	Reference	Prediction	WER_sentence	CER_sentence
asd02_008	de prinses heeft een euro	de prinses heeft een euro	0	0
asd12_002	en toen ging ze de bloemen hm hm water geven	en tek tek toen zei ze toen zei ze de bloemen om er water te geven	90	61.36
asd22_004	hij geeft de appel aan de indiaan	hij geeft de appel aan de indiaan	0	0
asd38_011	de ballerina wou water geven aan het gras	er was een heer wou een vaatschep uit de gras	100	70.73
asd38_012	en een kraan	vijna kraak	100	58.33
asd43_002	en ze geeft de bloemetje aan de ballerina de ballerina doet m in t haar	en ze geeft de bloemetje aan de ballerina de ballerina doet m in t haar	0	0

ASD only

Utterance	Reference	Prediction	WER_sentence	CER_sentence
asd02_008	de prinses heeft een euro	de prinses heeft een euro	0	0
asd02_009	de prinses heeft een ijsje gekocht en likt aan het ijsje	de prinses heeft een ijsje gekocht en likt aan het ijsje	0	0
asd12_002	en toen ging ze de bloemen hm hm water geven	hij kijkt toch in ze toezachsen toch in ze de bloemen om me eh water geven	110	84.09
asd19_004	alsjeblieft zei de cowboy	als je een blusje hebt kun je opnemen	200	96
asd19_009	en hij wil nog een ijsje	en hij wil nog een ijsje	0	0
asd38_010	ik weet niet wat te zeggen wat zeggen	ik ga me nu op de zeggen op de zeggen	87.5	45.95

TD+ADHD+ASD

Utterance	Reference	Prediction	WER_sentence	CER_sentence
asd02_004	de zuster plukt een bloem de zuster geeft de bloem aan de danseres danseres zet de bloem in zn haar	de zuster plukt een bloem de zuster geeft de bloem aan de danseres danseres zet de bloem in zn haar	0	0
asd10_007	piraat wil gaan voetballen	de piaar wil geen kopvang	100	57.69
asd12_002	en toen ging ze de bloemen hm hm water geven	en t t toekomst de zachtster toekomst de bloemen om m eh water te geven	100	81.82
asd19_001	een indiaan die pakt appels	een indiaan die pakt appels	0	0
asd22_003	hij valt van de steen en de appel hangt nog steeds in de boom een cowboy komt met een ladder aan de cowboy zet de ladder tegen de boom hij pakt de appel voor de indiaan	hij valt van de steen en de appel hangt nog steeds in de boom een cowboy komt met een ladder aan de cowboy zet de ladder tegen de boom hij pakt de appel voor de indiaan	0	0
asd38_010	ik weet niet wat te zeggen wat zeggen	ik ga me nu op de zegger op de zegger	112.5	51.35