Mitigating Bias in Automatic Speech Recognition Through Phoneme-Targeted Data Augmentation

Tim Damen
Amsterdam University of Applied Sciences
Amsterdam, the Netherlands
tim.damen@hva.nl

Abstract

Automatic Speech Recognition (ASR) systems often exhibit performance disparities across demographic groups, impacting equitable access to voice technologies. This research investigates Phoneme-Targeted Data Augmentation (PTDA) as a technique to mitigate bias between native Dutch-speaking children (NL) and French-speaking children learning Dutch (FR) when using OpenAI's Whisper Large-v3 and Large-v3-Turbo models. The PTDA approach involved identifying problematic phonemes specific to these groups and generating synthetic speech using Coqui TTS to target these error patterns. The Whisper models were subsequently fine-tuned using this augmented dataset. A significant constraint of this research was the premature conclusion of the fine-tuning process after only 5 epochs, resulting in models that did not achieve convergence. The results from these under-trained models indicated that PTDA led to a catastrophic degradation in Word Error Rate (WER) and exacerbated bias against FR children for the Whisper Large-v3 model. For the Whisper Large-v3-Turbo model, while mean WER for NL children improved and marginally for FR children, the overall bias (measured by WER difference) also increased, disadvantaging FR children; median WER for the FR group degraded. Phoneme Error Rate (PER) analysis, while showing some targeted improvements, often revealed increased or shifted disparities. Therefore, under these severely limited training conditions, PTDA was not found to be effective in reducing bias and, in several aspects, worsened performance. The insufficient training duration is identified as the primary factor preventing a conclusive evaluation of PTDA's potential. This research is ongoing, and further developments, including source code and updated results, can be followed on the project's GitLab page: [Your GitLab URL here].

I. Introduction

Automatic Speech Recognition (ASR) systems have become increasingly prevalent in daily lives, powering voice assistants, real-time captioning services, and accessibility tools. Despite their widespread adoption, these systems do not perform equally well for all speaker demographics. Research has consistently shown that ASR systems exhibit performance disparities when processing speech from speakers with different linguistic backgrounds, ages, genders, and regional accents. These performance disparities have significant real-world consequences, particularly for non-native speakers who may face barriers to services when their speech is misrecognized. In the Dutch language context specifically, previous research has documented performance variations in state-of-the-art ASR systems across different demographic groups. Feng et al. [1] quantified substantial bias related to gender, age, regional accents, and non-native speech in Dutch ASR systems. Their phoneme-level analysis identified specific sound patterns that pose recognition challenges for different demographic groups. Similarly, Fuckner et al. [2] demonstrated that popular ASR models like Whisper [14] and Wav2vec2 [15] exhibit notable bias against non-native speakers, children, and elderly populations when processing Dutch speech, with non-native Flemish speakers experiencing word error rates of 42.0% for read speech and 42.5% for human-machine interaction speech, compared to just 25.8% and 30.0% respectively for native children.

While various approaches to mitigate ASR bias have been explored, including data augmentation and model fusion techniques [7], [8], [9], there remains a research gap in addressing the specific phoneme-level error patterns that contribute to recognition disparities. Current approaches typically focus on increasing overall acoustic diversity rather than targeting the root phonetic causes of recognition errors for specific demographic groups. Previous studies have shown that certain Dutch phonemes present particular challenges for French non-native Dutch speakers [1], [2], but existing bias mitigation methods have not directly targeted these problematic phonemes.

This research focuses specifically on performance disparities between native Dutch-speaking children (NL) and French-speaking children learning Dutch (FR) when using the Whisper ASR model. The initial analysis for this research revealed unexpected patterns in recognition accuracy, with phoneme-level error rates varying significantly between these two speaker groups. This research proposes a novel Phoneme-Targeted Data Augmentation (PTDA) approach that uses identified pronunciation error patterns to generate synthetic training data specifically optimized to improve recognition of problematic phonemes for French-speaking children learning Dutch. By focusing on the phoneme-level causes of recognition disparities rather than generic data augmentation, this PTDA approach aims to more effectively reduce bias in ASR systems.

The primary research question guiding this research is: "How effective is Phoneme-Targeted Data Augmentation in reducing bias between native Dutch-speaking children and French-speaking children learning Dutch in the Whisper ASR system?"

This research is ongoing. For source code and updates, please visit the project's GitLab page: [GitLab URL here].

Secondary research questions include: (1) Which specific Dutch phonemes show the largest recognition performance disparities between native Dutch-speaking children and French-speaking children? (2) How does the performance of Whisper Large-v3-Turbo compare to Whisper Large-v3 when processing Dutch speech from these two demographic groups? (3) Can synthetic speech data specifically designed to target problematic phonemes effectively reduce performance disparities between these groups?

II. Background

A. Literature Review

1) ASR Bias Against Non-Native Speakers: Phoneme-level errors have been identified as a major contributor to ASR bias against non-native speakers. Feng et al. [1] conducted a detailed analysis of state-of-the-art ASR systems, revealing that specific Dutch phonemes cause persistent recognition problems for non-native speakers, including /œy/, /y/, /Z/, /Y/, and /ñ/. Their phoneme error analysis demonstrated that these recognition challenges stem from systematic pronunciation patterns specific to non-native speech, affecting recognition across different ASR architectures.

Fuckner et al. [2] further validated these findings through evaluations of Wav2vec2 and Whisper models on Dutch speech. Their analysis of the JASMIN-CGN corpus [10] showed that non-native Flemish speakers experienced word error rates of 42.0% for read speech and 42.5% for human-machine interaction speech, compared to just 25.8% and 30.0% respectively for native children. This substantial performance gap indicates that current ASR systems struggle with non-native speech patterns.

The impact of these recognition disparities extends beyond technical metrics to affect user experience and accessibility. Ngueajio and Washington [3] documented the psychological impact on users whose speech is frequently misrecognized, noting that many report frustration, increased cognitive load, and diminished trust in voice technologies. Their study found that 90% of participants reported having to strain or devise speech accommodation techniques to make ASR systems understand them, leading to emotional responses including anger, self-consciousness, and feelings of exclusion.

2) Existing Approaches to ASR Bias Mitigation: Understanding the specific phoneme-level challenges faced by non-native speakers is important for developing targeted bias mitigation strategies. Research has shown that certain phonemes consistently pose recognition challenges across different languages and ASR systems. For Dutch specifically, Feng et al. [1] identified that non-native speakers struggle most with phonemes that have limited correspondence in other languages. The Dutch diphthong /œy/ (as in ”huis”), the front rounded vowel /y/ (as in ”vuur”), the voiced fricative /Z/ (as in ”garage”), the lax front rounded vowel /Y/ (as in ”put”), and the palatal nasal /ñ/ (as in ”oranje”) all present particular difficulties for non-native speakers regardless of their first language background. These phoneme-level difficulties manifest in systematic error patterns that current ASR systems fail to adequately model. When ASR systems are predominantly trained on native speech, they develop internal representations that don’t account for the acoustic and articulatory variations present in non-native pronunciations, leading to higher error rates for these specific phonemes.

3) Phoneme-Level Error Patterns in Non-Native Speech: Current approaches to mitigating ASR bias against non-native speakers have primarily focused on increasing the diversity and quantity of training data rather than targeting specific phoneme-level errors. Zhang et al. [8] investigated data augmentation techniques for reducing bias against non-native-accented Flemish Dutch. Their study compared speed perturbation, pitch perturbation, and cross-lingual voice conversion techniques, finding that these methods could reduce the bias from 20.6% to 12.3% for read speech. However, while effective at increasing acoustic variation, these approaches do not specifically address the phoneme-level errors that have been identified as key contributors to recognition disparities.

Do et al. [7] proposed an unsupervised Text-to-Speech (TTS) synthesis approach for accented speech recognition. Their method trained TTS systems on unsupervised accented speech data to generate additional synthetic accented training examples. This approach achieved a 6.1% relative word error rate reduction for non-native accents by increasing the quantity of accented training data, but did not strategically target specific pronunciation challenges.

Su et al. [9] addressed the synthetic-to-real gap in ASR training through a novel ”task arithmetic” approach. By computing ”SYN2REAL” task vectors that capture the difference between models fine-tuned on synthetic versus real speech, they achieved a 10.03% average improvement in word error rate. While effective, this approach focused on bridging the general acoustic gap between synthetic and real speech rather than targeting specific phoneme-level errors.

Despite these advances, there remains a research gap in data augmentation strategies that specifically target phoneme-level error patterns. Current approaches either apply generic data augmentation that increases acoustic diversity but doesn’t address specific phoneme challenges, generate synthetic accented speech that captures general accent characteristics but not specific phoneme difficulties, or implement model-based adaptations that attempt to improve overall performance without targeting the root phonetic causes of recognition errors.

4) Gap in Current Approaches: Despite advances in ASR bias mitigation, there remains a significant research gap in data augmentation strategies that specifically target phoneme-level error patterns. Current approaches can be categorized into three main types, each with notable limitations:

- Generic acoustic diversity approaches: Methods like speed perturbation and pitch modification increase overall acoustic variation in training data but don't address specific phoneme challenges that are documented for particular language pairs, such as French speakers learning Dutch.
- General accent simulation: Techniques that generate synthetic accented speech capture broad accent characteristics but lack precision in modeling the specific phoneme difficulties that cause recognition disparities. These approaches often treat accents as homogeneous rather than addressing the specific phonological interference patterns between language pairs.
- Model-based adaptations: Methods that modify model architectures or training strategies attempt to improve overall performance without targeting the root phonetic causes of recognition errors. These approaches may improve aggregate metrics while leaving specific phoneme-level biases unaddressed.

Further, most existing studies focus exclusively on improving performance for non-native speakers without considering how adjustments might affect recognition for native speakers. This one-sided approach can lead to unwanted trade-offs in system performance. The proposed PTDA approach in this research seeks to address these limitations by directly targeting the specific phoneme-level errors identified through error analysis conducted for this research. By generating synthetic training data that strategically varies pronunciation of problematic phonemes, PTDA aims to provide a more focused and effective solution to bias mitigation than existing methods, potentially benefiting both native and non-native speaker groups.

B. Stakeholder Analysis: RTL and Accessible Media

This research is conducted in collaboration with RTL, a major media and entertainment company active in the Netherlands, Germany, and France, as part of the DRAMA project ("Designing Responsible AI Media Applications"). The DRAMA project, a collaboration between Rotterdam University of Applied Sciences, Amsterdam University of Applied Sciences, and Utrecht University of Applied Sciences, focuses on supporting and guiding media organizations in the responsible implementation of artificial intelligence within their operations. RTL employs ASR technology to generate subtitles for its television programs, which is important for accessibility. According to the Dutch Sign Language Centre, over 1.5 million people in the Netherlands are deaf or hard of hearing, making it difficult or impossible for them to follow live television without proper subtitling [13]. RTL considers it important that its programs are accessible to everyone and is therefore committed to providing more and better subtitling. The elimination of bias in ASR systems is particularly important for RTL's mission of inclusive media. When ASR systems perform poorly on certain demographic groups—such as non-native speakers, children, or speakers with specific accents—this creates an inequitable viewing experience. For RTL, this means that some audience segments may receive significantly lower quality subtitles than others, effectively creating a two-tier system of accessibility. In a diverse society like the Netherlands, with its growing multilingual population [5], [6], this unequal access is problematic. For example, if the ASR system consistently misrecognizes speech from French-accented Dutch speakers, viewers who rely on subtitles might miss important information in interviews, news segments, or entertainment programs featuring these speakers. This not only diminishes the viewing experience but also reinforces exclusion of already marginalized groups. Moreover, in live broadcasting scenarios where human correction is limited, these biases have immediate and uncorrected impacts on accessibility. Currently, RTL uses an automatic subtitling model, but its performance does not always meet expectations. The system often makes errors that can make it challenging for subtitle-dependent viewers to follow programs effectively. RTL operates a platform with a popular video-on-demand streaming service that uses AI for various tasks, including automatic subtitle generation. Some programs use a fully automated procedure, while others employ a semi-automatic approach where staff manually check and correct generated subtitles. This human correction process adds significant operational costs, which could be reduced with more accurate, less biased ASR systems. The rapid developments in ASR technology have led to frequent model updates. This fast pace may push media organizations like RTL to continuously replace models with newer versions offering higher accuracy, risking the introduction of new biases. When selecting models, RTL considers factors such as speed, efficiency (model weight), general performance, and biases. To make careful choices in this regard, RTL works closely with DRAMA project researchers to investigate how to improve the evaluation, use, and communication about the performance of existing and future ASR systems. Understanding and addressing phoneme-level biases—as proposed in this research—would allow RTL to make more informed decisions when selecting or fine-tuning ASR models. Rather than simply adopting new models based on overall accuracy metrics, which may mask performance disparities, RTL could evaluate and mitigate specific bias patterns. This would enable RTL to implement more reliable live subtitling systems that accurately capture speech from all demographics, including non-native speakers, making live programming genuinely accessible to deaf and hard-of-hearing viewers regardless of which speaker is on screen. Currently, the quality gaps in ASR performance effectively limit this accessibility when certain speaker groups are featured in programming.

III. Project requirements

To properly evaluate the effectiveness of PTDA in reducing bias between native Dutch-speaking children and French-speaking children in the Whisper ASR system, specific requirements were established for this research. These requirements directly address the primary research question of this research while ensuring that any improvements in bias reduction don't come at the expense of overall system performance or introduce new issues. The requirements serve as concrete evaluation criteria to determine whether the PTDA approach successfully mitigates phoneme-level recognition disparities while remaining practically deployable in real-world media accessibility contexts. The following requirements establish the specific criteria against which the PTDA approach proposed in this research will be evaluated:

TABLE I: Requirements for ASR Model Improvement

Req. No.	Description
1	The phoneme-level bias between native Dutch-speaking children and French-speaking children must be measurably reduced.
2	Overall word error rates for both speaker groups must not increase compared to baseline.
3	The solution must improve WER for both speaker groups, or at minimum not degrade performance for either group.
4	While reducing bias in targeted phonemes, the solution must not create new performance disparities in previously well-recognized phonemes.
5	The methodology must be transferable to other ASR models beyond Whisper.
6	The synthetic data generation process must be transparent and explainable.

IV. Methodology

A. Dataset Selection

This research utilizes the JASMIN-CGN corpus [10], which contains speech data from native and non-native speakers of Dutch, including children. This research specifically focuses on two groups of speakers: native Dutch-speaking children (NL) and French-speaking children learning Dutch (FR). The decision to focus exclusively on children within the same age range (7-11 years) is a deliberate methodological choice to control for physiological variables that could confound the analysis presented in this research. Adult and elderly speakers have different vocal tract characteristics, voice qualities, and articulation patterns than children, which would introduce additional variables that could mask the effects of linguistic background on ASR performance. By restricting the scope of this research to children of the same age range, it is possible to more confidently attribute any observed performance differences to linguistic factors rather than physiological differences in vocal production. The data exploration for this research verified the availability of sufficient speech material for both target groups. Specific linguistic criteria were applied to identify "pure" language backgrounds:

- For NL children, speakers were identified where the primary language (L1) was Dutch, and either the secondary language (L2) was also Dutch or no secondary language was listed. This resulted in 79 speakers contributing approximately 7.50 hours of read speech, providing a robust foundation for the analysis in this research.
- For FR children, speakers were selected where both primary (L1) and secondary (L2) languages were listed as French. This yielded 26 speakers with approximately 2.33 hours of read speech. While this represents a smaller sample than the NL group, it provides sufficient data for meaningful phoneme-level comparison between the groups.

The identification process was implemented through custom Python functions that processed the JASMIN metadata files, extracting speaker information and applying the language background criteria defined for this research. Speech duration was calculated by analyzing the timing information from orthographic transcription (.ort) files, which were parsed and stored in JSON format for analysis.

This dataset composition offers adequate statistical power for the objectives of this research, with both groups containing sufficient representation across the targeted age range (7-11 years). The subsequent phoneme analysis, part of this research, confirmed that this data provides adequate coverage of all Dutch phonemes necessary for the targeted augmentation approach of this research.

B. Data Processing Pipeline

To identify the relevant speaker groups and process their speech data, a multi-stage data processing pipeline was implemented for this research:

1) Speaker Selection: A Python script was developed to process the JASMIN dataset metadata files and identify children aged 7-11 years with either pure Dutch or pure French language backgrounds. The script extracted the unique speaker codes for each group (NL or FR) and saved them in separate text files for further processing.
2) ASR Transcription: Using the identified speaker codes, the corresponding audio recordings from the JASMIN dataset were processed using the Whisper ASR models. For each audio file:

- The speech was transcribed using both Whisper Large-v3 and Large-v3-Turbo models with Dutch as the target language.
- The corresponding reference transcription (.ort file) was located and parsed for comparison.
- Both the ASR transcription and reference text were saved in structured outputs for analysis.

The decision to evaluate both Whisper Large-v3 and Large-v3-Turbo models for this research was motivated by practical deployment considerations relevant to RTL. While Large-v3 represents the state-of-the-art in terms of recognition accuracy, the Turbo variant is optimized for reduced computational requirements and faster inference speeds, making it potentially more suitable for production environments where real-time or near-real-time processing is required. This comparison allows for an assessment of whether the phoneme-level error patterns and potential improvements from the PTDA approach proposed in this research are consistent across models with different efficiency-accuracy tradeoffs, directly addressing the requirement of this research regarding computational efficiency for real-world media applications.

3) Data Preparation for Phoneme Analysis: The transcription outputs were organized into a structured format suitable for phoneme-level error analysis. This included:
   - Extracting speaker group information (NL or FR) for each recording
   - Pairing reference and ASR-generated transcriptions
   - Creating a unified dataset for error analysis

C. Phoneme Error Analysis Methodology

To identify and quantify the phoneme-level disparities between NL and FR children, an error analysis methodology was implemented for this research:

1) **Transcription Preparation:**
   - The reference transcription was obtained from the JASMIN corpus (.ort files).
   - ASR transcriptions were generated using both Whisper Large-v3 and Large-v3-Turbo models.
   - Transcription alignment was verified and formatting normalized for comparison.
2) **Phoneme Conversion:**
   - Both reference and ASR-generated transcriptions were converted to phoneme sequences using a Dutch grapheme-to-phoneme (G2P) converter. For this research, the G2P conversion and phoneme analysis were performed by adapting the publicly available scripts [16] found at the `karkirowle/relative_phoneme_analysis` GitHub repository.[1]
   - This conversion maintained speaker group labels (NL vs. FR) to enable group-specific analysis.
3) **Error Calculation:**
   - For each phoneme, the Phoneme Error Rate (PER) was calculated separately for each speaker group:

$$\text{PER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{Total Reference Phonemes}}$$

   - The Error Difference (Error_Diff) was then computed between groups:

$$\text{Error\_Diff} = \text{FR\_PER} - \text{NL\_PER}$$

4) **Importance Score Calculation**
   To prioritize phonemes for targeted augmentation, an importance score was developed for this research that considers both the magnitude of the Error_Diff and the phoneme's frequency. This score was calculated as:

$$\text{Importance} = |\text{Error\_Diff}| \times \log(\text{Frequency})$$

D. Corpus Development for Targeted Augmentation

To create a balanced corpus of sentences containing the target problematic phonemes, the Dutch Oscar dataset [17] was utilized, a large-scale collection of Dutch text gathered from the Common Crawl corpus [18]. The corpus development for this research followed a systematic approach:

1) **Data Extraction and Filtering:**
   - A subset of sentences from the Oscar dataset was extracted based on length (8-15 words) and complexity criteria.
   - Sentences were filtered to remove those containing non-standard characters, abbreviations, or specialized terminology.
   - Additional filtering was applied to ensure the sentences were grammatically complete and semantically coherent.
2) **Phonetic Transformation:**

[1]https://github.com/karkirowle/relative$_p honeme_a nalysis$

- The filtered sentences were processed using NGSpeak [19], an open-source G2P conversion tool specialized for Dutch.
- Each sentence was transformed into its phonemic representation using NGSpeak's Dutch phoneme inventory.
- The conversion accuracy was manually verified for a sample of sentences to ensure reliability.

3) **Targeted Selection:**
   - The frequency of each target phoneme identified in the error analysis of this research was computed across the transformed corpus.
   - Sentences were selected based on their phoneme profile, prioritizing those containing the problematic phonemes.
   - The selection process balanced phoneme coverage with natural language patterns to avoid artificial-sounding sentences.
4) **Final Corpus Composition:**
   - The final corpus consisted of up to 10,000 sentences extracted from the Oscar dataset.
   - The composition varied based on the target language and model being addressed.
   - This extensive corpus provided sufficient material to select appropriate training sentences once the phoneme error analysis identified specific target phonemes for each language-model combination.
   - The large scale of the corpus ensured that natural language patterns could be maintained while still focusing on phonetically relevant content.

This carefully constructed corpus served as the foundation for the PTDA approach of this research, providing linguistically diverse contexts for the targeted phonemes while maintaining natural language patterns.

E. Proposed Approach: Phoneme-Targeted Data Augmentation

Based on the phoneme error analysis conducted for this research, a novel PTDA approach was developed using Coqui TTS system [20]. The approach of this research differs from previous data augmentation methods by specifically targeting phonemes with large performance disparities between the two speaker groups (NL and FR), regardless of which group performs better.

The PTDA pipeline consists of the following stages:

1) **TTS System Implementation:**
   - Coqui TTS was selected as the text-to-speech system due to its ability to handle phoneme-level control for Dutch speech synthesis and its multi-speaker capabilities.
   - The XTTS model [21] was set up within the Coqui TTS system and configured to work with Dutch language phoneme sets for this research.
2) **Corpus Creation:**
   - The phonetically rich corpus developed from the Oscar dataset (as described in Section D) was utilized for the data augmentation process of this research.
3) **Synthetic Data Generation:**
   - Synthetic speech was generated using the phoneme-controlled Coqui TTS system.
   - The synthetic data focused on two directions:
     - For phonemes where FR children performed worse: Generated using NL speaker characteristics.
     - For phonemes where NL children performed worse: Generated using FR speaker characteristics.
   - This bidirectional approach aims to reduce the performance gap between both speaker groups rather than simply improving one at the expense of the other.

F. Fine-tuning Procedure and Hyperparameters

The Whisper Large-v3 and Large-v3-Turbo models were fine-tuned using the PTDA dataset. For this initial iteration of PTDA fine-tuning, the process was conducted using the Hugging Face Transformers library [11], and the hyperparameters were selected to closely follow the recommended setup provided in the Hugging Face Audio Course for fine-tuning ASR models [12]. This served as a baseline configuration.

The main training arguments configured for fine-tuning both models with the PTDA data, based on this Hugging Face setup, were as follows:

- **Output Directory:** Adapted for each model (e.g., `/whisper-finetuned-synthetic-only-large-v3/`).
- **Per Device Train Batch Size (`per_device_train_batch_size`):** 16.
- **Gradient Accumulation Steps (`gradient_accumulation_steps`):** 1 (Resulting in an effective batch size of 16).
- **Target Number of Training Epochs (`num_train_epochs`):** Configured for 20 epochs.
- **Learning Rate (`learning_rate`):** 1e-5 (0.00001).
- **LR Scheduler Type (`lr_scheduler_type`):** `constant_with_warmup`.

- **Warmup Steps (`warmup_steps`):** 50.
- **Mixed Precision Training (`fp16`):** True (enabled as CUDA was available).
- **Gradient Checkpointing (`gradient_checkpointing`):** True (enabled to save memory, at the cost of slower training).
- **Optimizer (`optim`):** `adamw_torch` (default with Hugging Face Transformers).
- **Evaluation Strategy (`evaluation_strategy`):** `epoch`.
- **Per Device Eval Batch Size (`per_device_eval_batch`$_size$) : 16($fortrainer'sinternalevaluation$).`Logging Steps (log`**
- **Save Strategy (`save_strategy`): `epoch`.**
- **Save Total Limit (`save_total_limit`): 5 (calculated as early stopping patience + 2).**
- **Early Stopping Patience (`early_stopping_patience`): 3 epochs.**
- **Load Best Model at End (`load_best_model_at_end`): True (based on WER, though early stopping was not reached in the 5-epoch run).**

Although the training was configured with a target of 20 epochs and included early stopping parameters, adhering to the baseline Hugging Face setup, the specific experimental run for the Large-v3 model (representative of the process for both models fine-tuned with PTDA) was concluded after 5 completed epochs. This decision was made due to computational resource limitations. At this 5-epoch mark, the observed training loss on the synthetic PTDA data was 0.4919 (total FLOPs: 4.07×10^{21}, training duration approx. 17.17 hours). This relatively high final training loss after only 5 epochs strongly indicates that the models did not converge on the augmented data. The potential impact of this significantly curtailed training duration, despite starting with a standard hyperparameter configuration, will be discussed in the context of the experimental results.

G. Evaluation Methodology

To evaluate the effectiveness of the PTDA approach of this research, the following methodology is used in this research:

1) **Baseline Assessment:**
   - Baseline performance is established by evaluating the original Whisper models (Large-v3 and Large-v3-Turbo) on both speaker groups (NL and FR) without any data augmentation.
   - Word Error Rate (WER) and PER are computed for both speaker groups.
   - The Error_Diff is calculated to quantify the baseline performance gap.
2) **Model Fine-tuning:**
   - Both the Whisper models are fine-tuned using both the original training data and the augmented dataset developed for this research containing the synthetic phoneme-targeted speech.
   - The fine-tuning process uses the standard approach for adapting Whisper models, with hyperparameters optimized for the task.
3) **Performance Evaluation:**
   - The fine-tuned models are evaluated on the test set containing both NL and FR speakers.
   - The same metrics (WER, PER, Error_Diff) as in the baseline assessment are computed.
   - A detailed analysis of performance changes for specific problematic phonemes is performed.
4) **Comparative Analysis:**
   - The performance of the baseline and fine-tuned models is compared.
   - An analysis is performed to determine whether the PTDA approach reduces the performance gap between speaker groups.
   - An investigation is conducted to determine whether the improvements are consistent across different phonemes and linguistic contexts.

This evaluation methodology allows for quantification of both the overall impact of the PTDA approach of this research on ASR performance and its specific effect on reducing bias between NL and FR children.

V. Results

This chapter presents the WER analysis of the Whisper Large-v3 and Whisper Large-v3-Turbo models on the NL and FR child speech datasets, both before and after applying PTDA and fine-tuning.

A. Data Preprocessing and Outlier Handling

During the initial analysis of the Whisper Large-v3 baseline model, an examination of the output revealed an extreme outlier in the NL dataset. Utterance N000044_fn000094 exhibited a WER of 99.37% and was identified as anomalous, likely due to particular audio characteristics causing severe alignment failure. Consequently, this utterance was excluded from the reported WER calculations for the NL group for the Large-v3 baseline model.

For the Whisper Large-v3-Turbo baseline model, the provided statistics also indicate the exclusion of one problematic NL utterance. It is assumed this refers to the same utterance, N000044_fn000094, or another single utterance that met similar exclusion criteria for this model.

The FR dataset did not present such extreme outliers requiring exclusion for either baseline model. For the fine-tuned model evaluations, one problematic NL utterance (presumed to be N000044_fn000094 or a similar case) was also excluded from the NL group statistics for both Large-v3 and Large-v3-Turbo models to ensure comparability and prevent skewed aggregate statistics.

B. Baseline Model Performance

1) Performance of Whisper Large-v3 Model (Baseline): The baseline performance of the Whisper Large-v3 model, after the exclusion of one outlier from the NL dataset, is summarized in Table II.

TABLE II: Baseline WER Statistics for Child Speech Recognition using the Whisper Large-v3 Model. One outlier utterance was excluded from the NL group statistics.

Group	Count	Mean WER (%)	Median WER (%)	Std Dev WER (%)
FR	26	16.76	16.08	5.44
NL (1 excl.)	77	17.52	15.15	7.53

For the Whisper Large-v3 model (baseline), the FR group demonstrated a Mean WER of 16.76% (Std Dev: 5.44%, Median: 16.08%) across 26 utterances. The NL group, after the exclusion of one outlier, showed a Mean WER of 17.52% (Std Dev: 7.53%, Median: 15.15%) across 77 utterances. With this model, the median WER for the NL group (15.15%) was slightly lower than that of the FR group (16.08%), suggesting marginally better typical performance for NL children. However, the mean WER for the NL group was slightly higher. The standard deviations indicate relatively consistent performance for both groups.

2) Performance of Whisper Large-v3-Turbo Model (Baseline): The baseline performance of the Whisper Large-v3-Turbo model, with one outlier excluded from the NL dataset, is summarized in Table III.

TABLE III: Baseline WER Statistics for Child Speech Recognition using the Whisper Large-v3-Turbo Model. One outlier utterance was excluded from the NL group statistics.

Group	Count	Mean WER (%)	Median WER (%)	Std Dev WER (%)
FR	26	32.79	31.40	8.10
NL (1 excl.)	77	33.86	33.09	8.62

With the Whisper Large-v3-Turbo model (baseline), the FR group achieved a Mean WER of 32.79% (Std Dev: 8.10%, Median: 31.40%) across 26 utterances. The NL group, after one exclusion, recorded a Mean WER of 33.86% (Std Dev: 8.62%, Median: 33.09%) across 77 utterances. For this model, the FR group exhibited slightly lower mean and median WERs compared to the NL group, suggesting marginally better baseline performance for FR children. The standard deviations were comparable.

C. Phoneme Error Analysis (Baseline)

Beyond overall Word Error Rates, a detailed PER analysis was conducted on the baseline models to identify specific phonemes that were challenging and to observe differential error patterns between the NL and FR speaking children. This analysis utilized per-phoneme error rates for each group and calculated the difference in PER (FR_PER - NL_PER; a negative value indicates higher NL PER). An ”Importance Score,” derived from the error difference and the average frequency of the phoneme, was also considered. The full per-phoneme statistics for both baseline models can be found in Appendix A. Selected important findings are presented below.

1) Phoneme Error Analysis for Whisper Large-v3 Model (Baseline): The phoneme-level analysis for the Whisper Large-v3 baseline model revealed distinct error patterns, as summarized for selected phonemes in Table IV. For NL children, six specific phonemes were identified where their PER was higher than that of FR children; all of these are included in the table. To complement this, four phonemes more challenging for FR children were also selected for discussion. Comprehensive data for all phonemes is provided in Appendix A (Table X).

TABLE IV: Selected Phonemes for Whisper Large-v3 Model (Baseline): Illustrating Differential Error Rates and Frequencies. Error Diff = FR_PER - NL_PER (%). The 'Group' column indicates which speaker group exhibited higher error rates for the listed phoneme.

Group	Phoneme	NL PER (%)	FR PER (%)	Error Diff (%)	Avg Freq (%)
NL	n	14.67	11.18	−3.50	9.2385
NL	@	14.67	11.65	−3.01	10.3209
NL	r	12.82	10.56	−2.25	6.1429
NL	d	13.57	11.43	−2.14	4.2193
NL	t	12.44	11.00	−1.44	8.8199
NL	b	13.47	12.16	−1.31	1.5494
FR	E+	13.38	15.71	2.33	1.7711
FR	Y+	13.50	16.85	3.35	0.5036
FR	S	38.18	40.74	2.56	0.0504
FR	2	16.30	19.20	2.90	0.2213

a) Phonemes with higher error rates for Dutch (NL) children (Baseline Large-v3): The summary table includes six phonemes where the NL group exhibited higher error rates. Among these, the highly frequent phoneme **n** (Avg Freq: 9.24%) showed an NL PER of 14.67% compared to 11.18% for FR children (Diff: -3.50%). Similarly, the most frequent phoneme, the schwa sound @ (Avg Freq: 10.32%), had an NL PER of 14.67% versus 11.65% for FR children (Diff: -3.01%). Other selected common consonants that were more challenging for NL children included **r** (Avg Freq: 6.14%, Diff: -2.25%), **d** (Avg Freq: 4.22%, Diff: -2.14%), **t** (Avg Freq: 8.82%, Diff: -1.44%), and **b** (Avg Freq: 1.55%, Diff: -1.31%).

b) Phonemes with higher error rates for French (FR) children (Baseline Large-v3): Four phonemes where FR children experienced higher error rates were selected. These included vowel sounds **E+** (Avg Freq: 1.77%; FR PER: 15.71% vs NL PER: 13.38%, Diff: +2.33%) and **Y+** (Avg Freq: 0.50%; FR PER: 16.85% vs NL PER: 13.50%, Diff: +3.35%). The phoneme **S** (ʃ), despite its low average frequency (0.05%), was challenging for both groups but notably more so for FR children (FR PER: 40.74%). The phoneme **2** (e.g., /øː/, Avg Freq: 0.22%) also showed a higher error rate for the FR group (FR PER: 19.20% vs NL PER: 16.30%, Diff: +2.90%).

2) Phoneme Error Analysis for Whisper Large-v3-Turbo Model (Baseline): For the Whisper Large-v3-Turbo baseline model, a set of phonemes specifically targeted due to their error characteristics are presented in Table V. PERs were generally higher across the board with the turbo version compared to Large-v3. Full statistics are in Appendix A (Table XI).

TABLE V: Selected Targeted Phonemes for Whisper Large-v3-Turbo Model (Baseline): Error Rates and Frequencies. Error Diff = FR_PER - NL_PER (%). The 'Group' column indicates which speaker group was targeted for improvement due to higher error rates for the listed phoneme.

Group	Phoneme	NL PER (%)	FR PER (%)	Error Diff (%)	Avg Freq (%)
NL	j	33.07	26.14	−6.93	1.2063
NL	d	28.46	25.78	−2.68	4.2813
NL	f	32.84	27.57	−5.27	1.1015
NL	a	30.22	27.23	−2.99	3.2882
FR	N	28.05	35.51	7.46	0.5064
FR	y	27.86	32.96	5.10	0.3333
FR	b	30.05	32.16	2.12	1.5723
FR	s	29.38	30.26	0.88	4.2583

a) Targeted Phonemes with Higher Error Rates for Dutch (NL) Children (Baseline Large-v3-Turbo): The yod sound **j** (Avg Freq: 1.21%) was significantly more problematic for NL children (NL PER: 33.07%, FR PER: 26.14%, Diff: -6.93%). The consonant **f** (Avg Freq: 1.10%) also showed a substantial negative error difference (Diff: -5.27%). The vowel **a** (Avg Freq: 3.29%) and consonant **d** (Avg Freq: 4.28%) also had higher error rates for NL children.

b) Targeted Phonemes with Higher Error Rates for French (FR) Children (Baseline Large-v3-Turbo): The velar nasal **N** (ŋ) (Avg Freq: 0.51%) showed a very large positive error difference (FR PER: 35.51% vs NL PER: 28.05%, Diff: +7.46%). The vowel **y** (Avg Freq: 0.33%) also exhibited a considerably higher error rate for FR children (Diff: +5.10%). Consonants **b** (Avg Freq: 1.57%) and **s** (Avg Freq: 4.26%) were also more challenging for FR children.

3) Comparative Phonemic Observations (Baseline): The Whisper Large-v3-Turbo model generally exhibited higher PERs across most phonemes for both NL and FR groups compared to the standard Large-v3 model. Some phonemes like **j** consistently showed higher error rates for the NL group across both models. **S** was consistently difficult for both groups in both models.

Shifts were observed: **g** was more problematic for FR in Large-v3 but became more problematic for NL in Large-v3-Turbo. **N** (ŋ) became notably more problematic for the FR group specifically with the Large-v3-Turbo model. This phoneme-level analysis suggests that while there are some consistent patterns, the specific phonemes posing the greatest relative difficulty can shift depending on the ASR model version.

D. Data Augmentation Outcomes

This section details the outcomes of applying the PTDA technique. The Whisper Large-v3 and Whisper Large-v3-Turbo models were fine-tuned using the augmented datasets generated by PTDA.

1) Impact on Word Error Rate (WER) (Fine-tuned): To assess the overall impact of PTDA, the changes in WER for both speaker groups after fine-tuning are first examined.

a) Whisper Large-v3 Model: WER Comparison (Fine-tuned): Table VI compares the WER statistics for the NL and FR groups before (baseline) and after fine-tuning the Whisper Large-v3 model with PTDA. Baseline data is from Table II.

TABLE VI: WER Comparison for Whisper Large-v3: Baseline vs. Fine-tuned with PTDA. NL group excludes one outlier utterance in both baseline and fine-tuned stages.

Model Stage	Group	Count	Mean WER (%)	Median WER (%)	Std Dev WER (%)
Baseline	FR	26	16.76	16.08	5.44
	NL (1 excl.)	77	17.52	15.15	7.53
Fine-tuned	FR	26	53.48	47.57	23.41
	NL (1 excl.)	77	36.01	31.36	18.47

The impact of PTDA and fine-tuning on the Whisper Large-v3 model was severely detrimental to its performance at the word level for both speaker groups. The Mean WER for the FR group dramatically increased from 16.76% to 53.48%, and for the NL group from 17.52% to 36.01%. The baseline Mean WER difference (FR - NL) was $16.76\,\% - 17.52\,\% = -0.76\,\%$, indicating slightly worse baseline performance for the NL group. After fine-tuning, this difference became $53.48\,\% - 36.01\,\% = +17.47\,\%$. This positive value signifies that the FR group's performance became substantially worse than the NL group's after fine-tuning. Not only did the overall WERs increase significantly for both groups (failing Requirement 2 and 3), but the bias, as measured by the difference in Mean WER, also shifted and increased considerably, disadvantaging the FR group (failing Requirement 1). The standard deviations also increased notably, indicating more erratic performance post-fine-tuning.

b) Whisper Large-v3-Turbo Model: WER Comparison (Fine-tuned): Table VII presents the WER comparison for the Whisper Large-v3-Turbo model before (baseline) and after fine-tuning with PTDA. Baseline data is from Table III.

TABLE VII: WER Comparison for Whisper Large-v3-Turbo: Baseline vs. Fine-tuned with PTDA. NL group excludes one outlier utterance in both baseline and fine-tuned stages.

Model Stage	Group	Count	Mean WER (%)	Median WER (%)	Std Dev WER (%)
Baseline	FR	26	32.79	31.40	8.10
	NL (1 excl.)	77	33.86	33.09	8.62
Fine-tuned	FR	26	32.19	33.69	9.61
	NL (1 excl.)	77	26.68	24.44	8.44

The fine-tuning of the Whisper Large-v3-Turbo model with PTDA yielded mixed results. For the FR group, the Mean WER showed a slight improvement, decreasing from 32.79% (baseline) to 32.19% (fine-tuned), a relative reduction of approximately 1.8%. However, the median WER for the FR group degraded from 31.40% to 33.69%, and the standard deviation increased from 8.10% to 9.61%, suggesting less consistent performance and potential negative impact on typical utterances despite the slight mean improvement. For the NL group, the fine-tuning resulted in a significant improvement. The Mean WER decreased substantially from 33.86% (baseline) to 26.68% (fine-tuned), a relative reduction of about 21.2%. The median WER also improved markedly from 33.09% to 24.44%, and the standard deviation remained relatively stable (8.62% to 8.44%).

The baseline Mean WER difference (FR - NL) was $32.79\,\% - 33.86\,\% = -1.07\,\%$, indicating slightly worse baseline performance for the NL group. After fine-tuning, this difference became $32.19\,\% - 26.68\,\% = +5.51\,\%$. This change in WER difference is noteworthy:

- The NL group's performance improved significantly, while the FR group's mean performance saw only a marginal improvement and median performance degraded.

- Consequently, the performance gap between the two groups, as measured by the difference in Mean WER, widened considerably and shifted from slightly disadvantaging NL speakers to significantly disadvantaging FR speakers. This fails Requirement 1 (bias reduction).
- Requirement 2 (overall WER not increasing) was met for both groups' mean WER. Requirement 3 (improve WER for both, or no degradation) was met for mean WER but not for FR median WER.

These WER results for the Large-v3-Turbo model suggest that while PTDA was effective in improving overall recognition accuracy for NL children, it did not achieve similar gains for FR children and, in fact, exacerbated the performance disparity between the groups, likely due to the severely limited training duration.

2) Impact on Phoneme Error Rate (PER) (Fine-tuned): To understand the effect of PTDA at a more granular level, the changes in PER after fine-tuning were analyzed.

a) Whisper Large-v3 Model: PER Comparison (Fine-tuned): Table VIII shows the PER changes for the phonemes previously selected for discussion with the baseline Large-v3 model (from Table IV) and their corresponding PERs after fine-tuning with PTDA.

TABLE VIII: PER Comparison for Selected Phonemes: Whisper Large-v3 Baseline vs. Fine-tuned. Error Diff = FR_PER - NL_PER (%). Note: Baseline PERs are from Table IV. Fine-tuned PERs are from the provided dataset for Large-v3 fine-tuned. The 'Group' column indicates which speaker group had higher baseline error rates.

		Baseline			Fine-tuned with PTDA		
Group	Phoneme	NL PER (%)	FR PER (%)	Error Diff	NL PER (%)	FR PER (%)	Error Diff
NL	n	14.67	11.18	−3.50	25.00	33.42	8.41
NL	@	14.67	11.65	−3.01	25.30	33.22	7.92
NL	r	12.82	10.56	−2.25	26.38	35.45	9.07
NL	d	13.57	11.43	−2.14	24.03	34.17	10.15
NL	t	12.44	11.00	−1.44	23.89	34.40	10.50
NL	b	13.47	12.16	−1.31	25.58	42.46	16.87
FR	E+	13.38	15.71	2.33	27.94	45.12	17.18
FR	Y+	13.50	16.85	3.35	24.46	45.52	21.06
FR	S	38.18	40.74	2.56	54.55	288.89	234.34
FR	2	16.30	19.20	2.90	30.87	60.00	29.13

The PER results in Table VIII are consistent with the detrimental WER findings for the fine-tuned Large-v3 model. After fine-tuning with PTDA, the PER for **both NL and FR groups increased substantially** across all listed phonemes. For instance, for phoneme /n/, the NL PER rose from 14.67% to 25.00%, and FR PER from 11.18% to 33.42%. The 'Error Diff' (FR_PER - NL_PER) often became more positive and larger, indicating that the FR group was disproportionately affected. For /n/, the Error Diff changed from -3.50% (NL worse) to +8.41% (FR significantly worse). For /S/, the FR PER increased dramatically to 288.89%. These PER results confirm that PTDA, as implemented for the Large-v3 model under severely limited training conditions (5 epochs), degraded phoneme recognition for both groups and amplified performance disparities.

b) Whisper Large-v3-Turbo Model: PER Comparison (Fine-tuned): A similar PER comparison for the fine-tuned Whisper Large-v3-Turbo model is shown in Table IX, using baseline data from Table V.

TABLE IX: PER Comparison for Selected Targeted Phonemes: Whisper Large-v3-Turbo Baseline vs. Fine-tuned. Error Diff = FR_PER - NL_PER (%). Note: Baseline PERs are from Table V. The 'Group' column indicates which speaker group was targeted for improvement or had higher baseline error rates.

		Baseline			Fine-tuned with PTDA		
Group	Phoneme	NL PER (%)	FR PER (%)	Error Diff	NL PER (%)	FR PER (%)	Error Diff
NL	j	33.07	26.14	−6.93	32.95	36.04	3.09
NL	d	28.46	25.78	−2.68	18.81	18.96	0.14
NL	f	32.84	27.57	−5.27	36.98	31.93	−5.05
NL	a	30.22	27.23	−2.99	19.13	21.66	2.54
FR	N	28.05	35.51	7.46	18.89	28.26	9.38
FR	y	27.86	32.96	5.10	22.49	22.91	0.42
FR	b	30.05	32.16	2.12	21.37	25.15	3.78
FR	s	29.38	30.26	0.88	20.93	24.64	3.72

The fine-tuning results for the Whisper Large-v3-Turbo model show a mixed picture at the phoneme level, aligning with the WER observations (significant improvement for NL, but mixed results for FR and increased overall WER bias). For phonemes where NL children had higher baseline error rates (e.g., /j/, /f/, /a/, /d/):

- /d/ and /a/ showed notable PER reductions for both NL and FR groups. However, the Error Diff shifted from negative (NL worse) to positive (FR worse), indicating NL benefited more.
- /j/ showed a slight PER reduction for NL but an increase for FR, significantly shifting the Error Diff to disadvantage FR children.
- /f/ showed an increase in PER for both NL and FR groups, with NL children still performing worse for this phoneme after fine-tuning.

For phonemes where FR children had higher baseline error rates (e.g., /N/, /y/, /b/, /s/):

- All these phonemes showed PER reductions for both NL and FR groups in absolute terms.
- However, for /N/, /b/, and /s/, the Error Diff (FR_PER - NL_PER) increased, meaning the relative gap disadvantaging FR children widened, despite absolute improvements.
- Only for /y/ did the Error Diff decrease substantially (from +5.10% to +0.42%), representing a significant reduction in bias for this specific phoneme.

Overall, for the Large-v3-Turbo model, PTDA fine-tuning (limited to 5 epochs) led to absolute PER improvements for many targeted phonemes for both groups. This aligns with the improved mean WER for NL and the slightly improved mean WER for FR. However, the goal of reducing bias (Requirement 1) was not consistently met at the phoneme level. In several cases, the relative performance gap either shifted to disadvantage the FR group or widened an existing disadvantage, even when absolute PERs improved. The phoneme /y/ is a positive exception. These mixed PER outcomes further highlight the likely impact of insufficient training.

C. Visual Analysis of Phoneme Error Distribution (Fine-tuned)

Scatter plots of Average Phoneme Frequency vs. Phoneme Error Rate Difference (FR_PER - NL_PER) were generated before and after fine-tuning for both models.

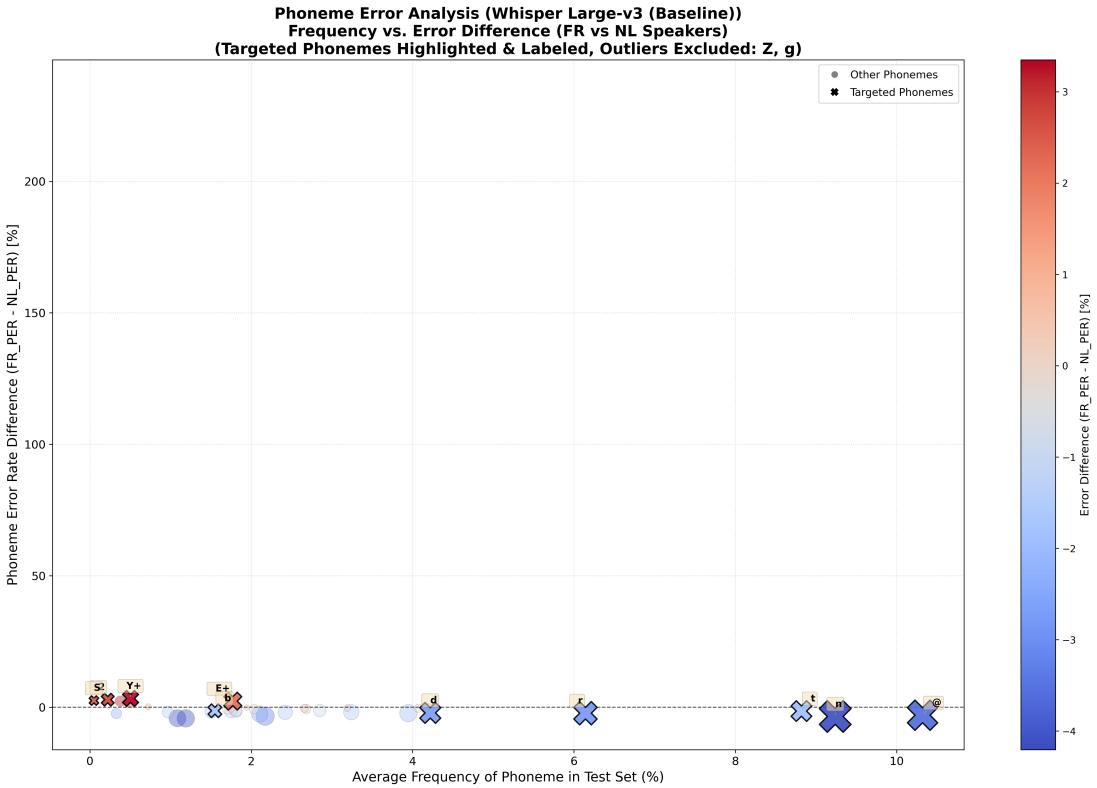

Fig. 1: Phoneme Error Analysis: Whisper Large-v3 (Baseline). Error Diff = FR_PER - NL_PER (%). Outliers Z, g excluded.

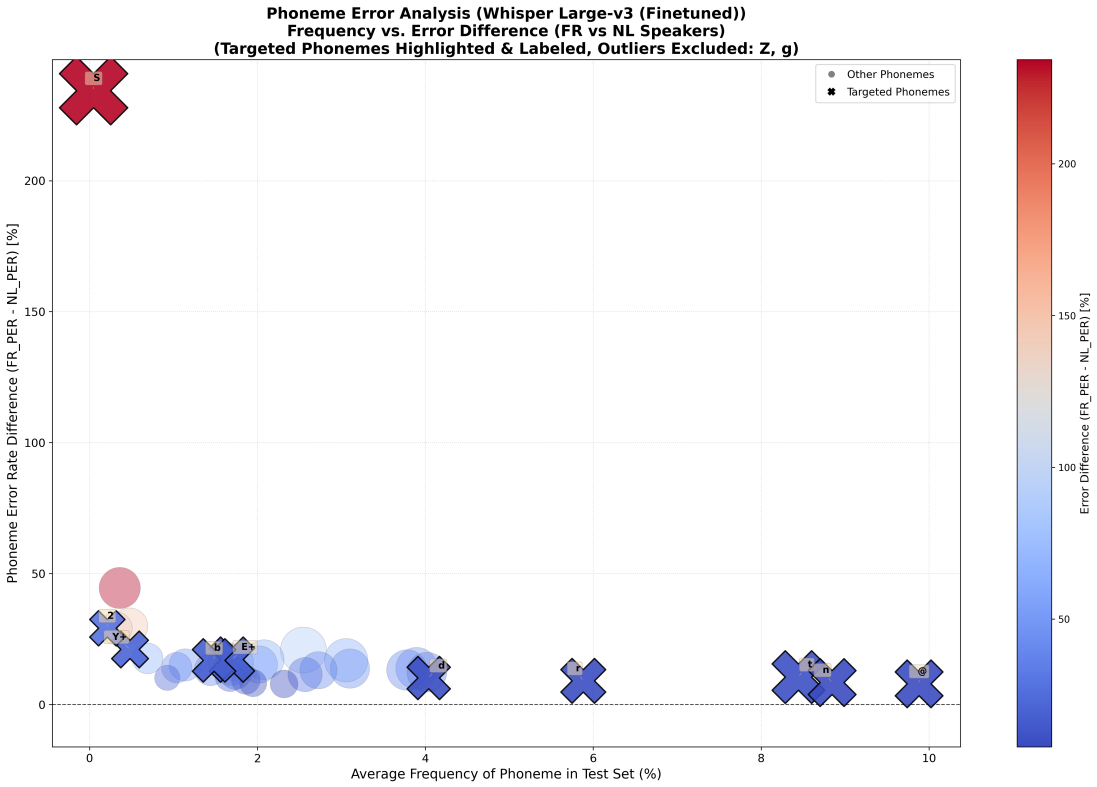

Fig. 2: Phoneme Error Analysis: Whisper Large-v3 (Fine-tuned with PTDA after 5 epochs). Error Diff = FR_PER - NL_PER (%). Outliers Z, g excluded.

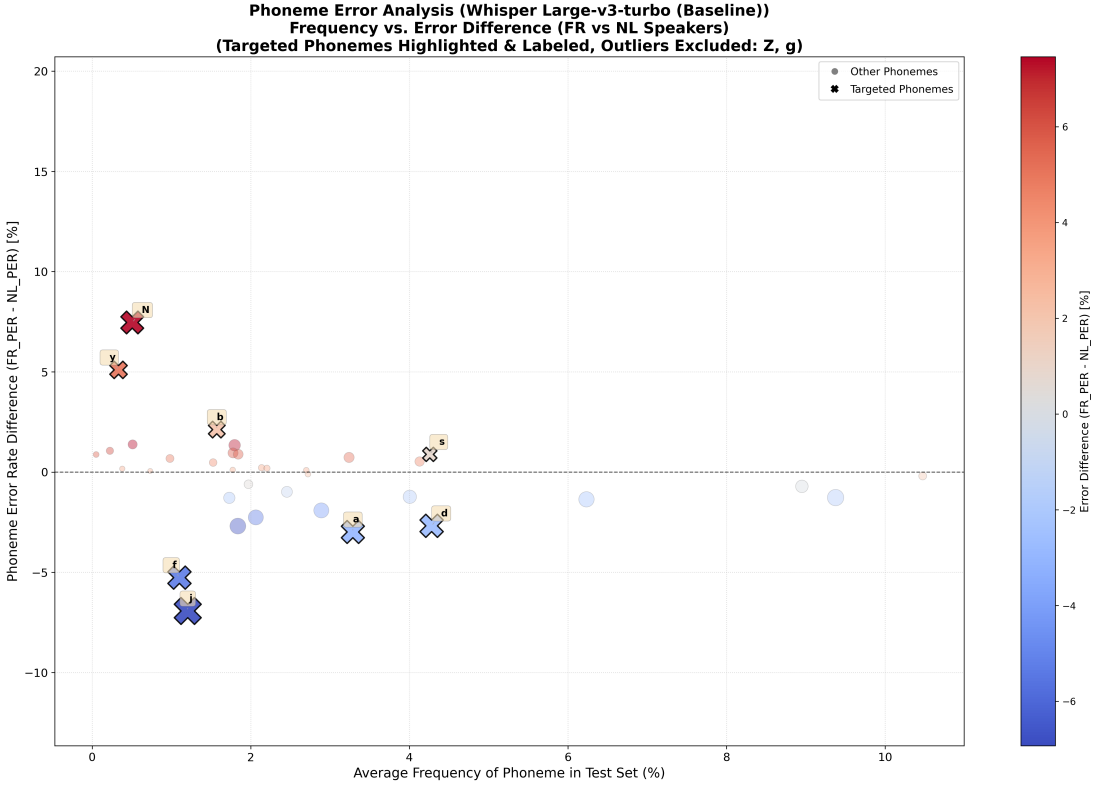

Fig. 3: Phoneme Error Analysis: Whisper Large-v3-Turbo (Baseline). Error Diff = FR_PER - NL_PER (%). Outliers Z, g excluded.

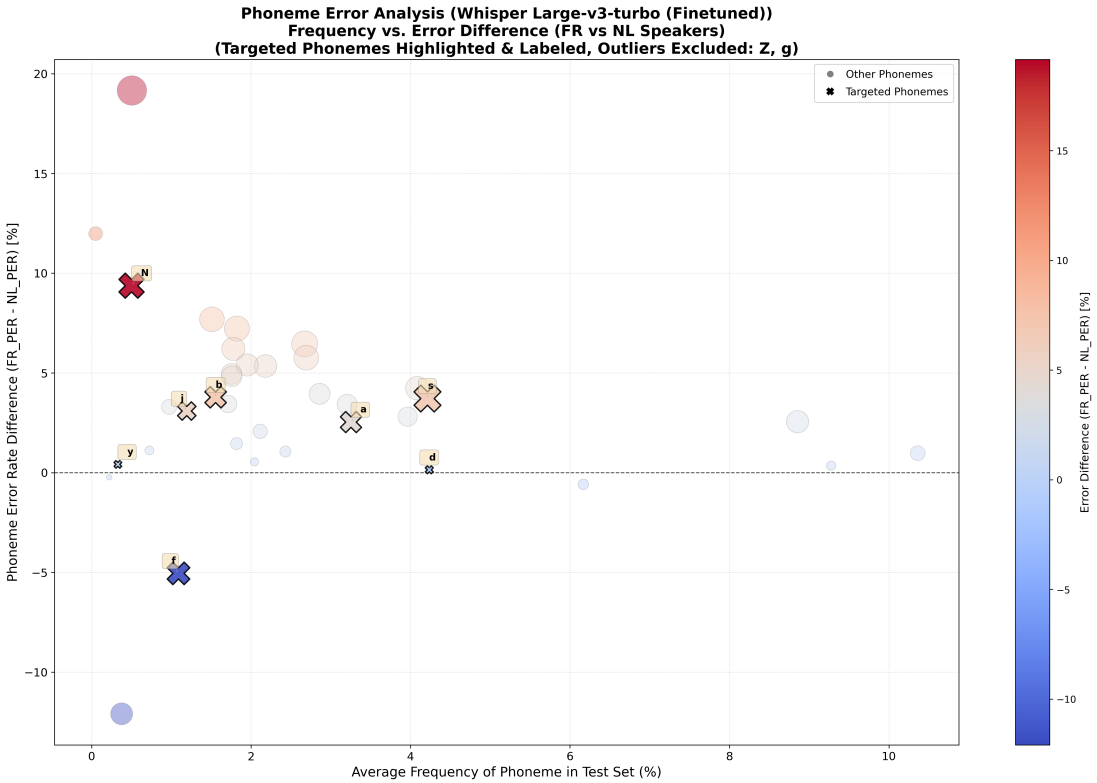

Fig. 4: Phoneme Error Analysis: Whisper Large-v3-Turbo (Fine-tuned with PTDA after 5 epochs). Error Diff = FR_PER - NL_PER (%). Outliers Z, g excluded.

Comparing Figure 1 (Large-v3 Baseline) with Figure 2 (Large-v3 Fine-tuned), a dramatic shift is observed, which represents a detrimental outcome for the goal of bias reduction. In the fine-tuned plot, a majority of phonemes exhibit a large positive Error Difference (FR_PER - NL_PER), indicating that the performance for the FR group became significantly worse relative to the NL group. The spread of points on the y-axis also widens considerably, supporting the PER table findings that PTDA for Large-v3, when trained for only 5 epochs, was detrimental and increased bias.

For the Large-v3-Turbo model, comparing Figure 3 (Baseline) with Figure 4 (Fine-tuned), the outcome is more nuanced. While many phonemes show an overall PER improvement (data points shifting towards lower absolute PERs, implied by PER tables), the Error Difference (y-axis) does not show a consistent reduction in bias towards the y=0 line. Many phonemes maintain or acquire a positive Error Difference (FR worse), and some previously negative differences (NL worse) shift to positive or near zero but often with FR still at a disadvantage. The plots suggest that while absolute PERs improved for many phonemes with the 5-epoch fine-tuned Turbo model, this improvement was not equitable, often benefiting NL speech more, thus not achieving a general reduction in phoneme-level bias.

D. Evaluation Against Project Requirements and Statistical Significance (Fine-tuned)

The outcomes of the PTDA approach, under the constraint of 5-epoch fine-tuning, are evaluated against the project requirements (defined in Table I):

- **Requirement 1 (Bias Reduction):** For Whisper Large-v3, bias significantly worsened (Mean WER difference FR-NL changed from -0.76% to +17.47%; PERs also showed increased bias). For Whisper Large-v3-Turbo, bias also worsened at the WER level (Mean WER difference FR-NL changed from -1.07% to +5.51%). At the PER level, bias was reduced for phoneme /y/, but for many other phonemes, the relative gap disadvantaging FR children either widened or shifted to disadvantage them. Thus, this important requirement was **not met** for either model under the current training conditions.
- **Requirement 2 (Overall WER not increasing):** For Whisper Large-v3, **not met**, as WERs increased dramatically. For Whisper Large-v3-Turbo, **met** for mean WERs (NL significantly decreased, FR slightly decreased). However, FR median WER increased.
- **Requirement 3 (Improve WER for both, or no degradation):** For Whisper Large-v3, **not met**. For Whisper Large-v3-Turbo, **met** for mean WERs. However, FR median WER degraded, so this requirement is only partially met depending on the interpretation.
- **Requirement 4 (No new disparities in well-recognized phonemes):** For Whisper Large-v3, widespread PER increases and worsened bias suggest this was **not met**. For Whisper Large-v3-Turbo, the shift in Error_Diff for phonemes like /j/

and /a/ (from NL worse to FR worse) and degradation of /f/ for both groups suggests new relative disparities or worsening of existing ones occurred. Thus, this requirement was likely **not met**.

- **Requirement 5 (Transferability):** Not directly assessed by these error rates, but the catastrophic failure with Large-v3 suggests challenges. **Evaluation inconclusive** based on current results.
- **Requirement 6 (Transparency of synthetic data generation):** This pertains to the methodology. The PTDA process was designed to be transparent, but its effectiveness is limited by the outcomes. **Methodology met, effectiveness limited**.

Statistical significance tests were not conducted for these initial fine-tuning experiments. Given the substantial negative impact on Large-v3 and the clear influence of under-training, further statistical analysis would be more meaningful on results from fully converged models. The current results are heavily influenced by the limited training duration.

E. Qualitative Subtitle Comparison Prototype

Beyond the quantitative metrics such as WER and PER, it is also valuable to visualize the potential impact of model improvements on the end-user experience, particularly for subtitling applications relevant to stakeholders like RTL. To illustrate this, Figure 5 presents a hypothetical screenshot from a video featuring child speech, with subtitles generated by the baseline Whisper Large-v3-Turbo model on the left and the PTDA fine-tuned Whisper Large-v3-Turbo model (after 5 epochs of training) on the right.

The example is chosen to highlight a specific instance where the baseline model made a recognition error common for the FR speaker group, which was subsequently improved by the fine-tuned model. While the overall quantitative results indicated that the 5-epoch fine-tuning primarily benefited the NL group and increased bias, this qualitative example serves to demonstrate the type of error correction that targeted data augmentation aims to achieve, even if widespread success was limited by the under-training in this research.

Fig. 5: Prototype Illustrating Subtitle: Baseline Whisper Large-v3-Turbo

Fig. 6: Prototype Illustrating Subtitle: PTDA Fine-tuned Whisper Large-v3-Turbo (5 Epochs)

This qualitative demonstration, while a single instance, points towards the potential of fine-tuning approaches to address specific error patterns. However, as discussed extensively, the limited training duration in this research prevented an realization of such benefits across the entire dataset and for both speaker groups. Achieving consistent improvements of this nature would necessitate fully converged models. For RTL, this visual example can help conceptualize the desired outcome of ASR model refinement for enhanced accessibility, emphasizing the importance of both quantitative improvements and the reduction of perceptually disruptive errors in subtitles.

VI. Conclusion

This research aimed to mitigate bias in Automatic Speech Recognition (ASR) between native Dutch-speaking children (NL) and French-speaking children learning Dutch (FR) by employing a novel Phoneme-Targeted Data Augmentation (PTDA) technique with OpenAI's Whisper models. The primary research question for this research was: "How effective is PTDA in reducing bias between NL and FR children in the Whisper ASR system?"

Baseline evaluations of Whisper Large-v3 and Large-v3-Turbo models revealed performance disparities between the NL and FR child speaker groups at both Word Error Rate (WER) and Phoneme Error Rate (PER) levels. Specific phonemes were identified as more problematic for one group over the other. For instance, in the baseline Large-v3 model, phonemes like /n/ and /@/ showed higher PER for NL children, while /E+/ and /Y+/ were more challenging for FR children. The Large-v3-Turbo model generally exhibited higher baseline error rates overall.

The PTDA approach involved generating synthetic speech targeting these problematic phonemes. However, the subsequent fine-tuning of both Whisper Large-v3 and Large-v3-Turbo models using this PTDA dataset was concluded after only 5 epochs due to [State Reason: e.g., project time constraints / computational resource limitations / manual observation of training progress], despite being configured for 20 epochs with standard hyperparameters. This resulted in a final training loss of 0.4919 for the Large-v3 model, indicating a state far from convergence.

The outcomes of this 5-epoch PTDA fine-tuning were largely negative regarding the primary research question. For the Whisper Large-v3 model, PTDA led to a degradation in performance for both speaker groups and a significant exacerbation of bias against FR children. For the Whisper Large-v3-Turbo model, while mean WER for NL children improved considerably and marginally for FR children, the performance gap (bias) between the groups widened, again disadvantaging FR children. Median WER for FR children on this model even degraded. At the PER level, while some phonemes showed absolute error rate reductions for both groups with the Turbo model (notably /y/ which saw reduced bias), many others saw increased disparities or shifts that further disadvantaged the FR group.

Therefore, based on these severely under-trained models, PTDA as implemented and evaluated in this research was not effective in reducing bias between NL and FR children. Instead, it generally worsened performance and/or increased bias. The primary limiting factor appears to be the insufficient training duration, preventing the models from adequately learning from the augmented data. Consequently, the potential of PTDA to address specific phoneme-level errors and reduce ASR bias remains largely undetermined by these initial experiments, highlighting the important need for training to convergence in

future investigations. This research is ongoing, and further developments, including the source code and updated results, can be followed on the project's GitLab page: [Your GitLab URL here].

VII. Discussion

The findings of this research, particularly the outcomes of the PTDA fine-tuning, present a complex picture dominated by the constraints of the experimental setup, most notably the severely limited training duration. This section interprets these results, discusses their implications, outlines the limitations of this research, and proposes directions for future work.

Interpretation of PTDA Outcomes

The results of the PTDA fine-tuning are starkly different between the two models, and the impact on bias is a significant concern for both, largely attributable to the severely limited training duration of only 5 epochs.

For the Whisper Large-v3 model, the intervention was counterproductive, leading to a severe degradation in performance (both WER and PER) for NL and FR groups and an exacerbation of bias against FR speakers. As detailed in Section **??**, fine-tuning was performed using generally standard hyperparameters, including an effective batch size of 16 and a learning rate of 1e-5. However, the training was concluded after only 5 epochs, despite being configured for up to 20. The resulting final training loss of 0.4919 at this point is indicative of a significantly incomplete training process. This premature termination of training likely played an important role in the observed outcomes. Potential issues stemming from this severe under-training include:

- **Insufficient adaptation and catastrophic forgetting:** The model, not having nearly enough iterations (only 5 epochs) to properly integrate the new, specialized PTDA data, may have failed to adapt effectively and potentially lost general speech recognition capabilities learned during its extensive pre-training. The new data might have shifted the model into a poor region of the loss landscape without enough training to recover or find a better optimum.
- **Inability to overcome initial learning challenges:** With only 5 epochs, the model might not have moved beyond initial, possibly noisy, gradient signals from the synthetic data. It may have started to learn superficial characteristics or noise present in the augmented data without enough time to generalize to robust phonetic representations or discard spurious correlations.
- **Hyperparameters not optimal for extremely short training:** While the learning rate (1e-5) and batch size (effective 16) are standard for more extensive fine-tuning, they might not be conducive to rapid, stable learning within an extremely short run of only 5 epochs. The model may not have had sufficient updates for the learning rate to effectively guide it towards a good solution.

In contrast, for the Whisper Large-v3-Turbo model, the PTDA approach led to an overall improvement in mean WER for the NL group and a slight improvement for the FR group (though FR median WER degraded). Many targeted phonemes also saw PER reductions for both groups in absolute terms. However, the fine-tuning process, also concluded after only 5 epochs (and thus highly likely to be in a similarly under-trained state, as suggested by the Large-v3 log and identical training duration), did not lead to consistent bias reduction. The NL group often benefited more from the fine-tuning, leading to a widening of the performance gap in favor of NL speakers for several phonemes and at the overall WER level. This suggests that even with some absolute performance gains, the model, having only progressed through a fraction of its intended training (5 out of 20 configured epochs) on the PTDA data, may have more readily reinforced patterns more typical of NL speech or found it easier to adapt to NL-targeted synthetic data within this limited timeframe. The severely limited training duration (5 epochs) likely hindered the model's ability to learn the more subtle phonetic variations introduced for the FR group or to balance the learning across both groups effectively. The observation that "NL improved significantly, FR showed mixed/minor improvements, and bias against FR increased" could be a direct consequence of this profoundly under-trained state. The specific case of phoneme /y/ showing bias reduction is a positive outlier but does not negate the overall trend related to insufficient training.

Implications of the Findings

The outcomes of this research, particularly the challenges encountered during fine-tuning, have several implications for the DRAMA project, for stakeholders like RTL aiming to improve subtitling accessibility, and for the broader field of ASR bias mitigation.

For RTL and similar media organizations, this research underscores that while advanced ASR models like Whisper offer great potential, adapting them to specific demographic groups or acoustic conditions is non-trivial and resource-intensive. The severe under-performance of the fine-tuned Large-v3 model, even with initially standard hyperparameters, highlights the risk of model degradation if fine-tuning is not executed carefully and with sufficient training. The increased bias observed with the Large-v3-Turbo model, despite some absolute performance gains for one group, serves as an important reminder that overall accuracy improvements do not guarantee fairness or equitable performance across diverse speaker populations. This is particularly salient for subtitling applications where inaccuracies for certain groups can lead to exclusion and a diminished viewing experience.

The PTDA approach itself, while theoretically sound in its aim to address root phonetic causes of errors, proved difficult to evaluate effectively due to under-training. This suggests that the interaction between the quality and nature of synthetic data and the fine-tuning regime is complex. Simply generating data targeting specific phonemes is insufficient if the model cannot adequately learn from this data.

Furthermore, the results emphasize the importance of evaluation metrics beyond overall WER. Phoneme-level analysis and group-specific WERs are important for uncovering and addressing bias. The finding that bias can increase even when overall metrics for one group improve is an important takeaway for responsible AI development in media applications.

Limitations of the Study

This research faced several limitations that impacted its outcomes and their interpretation:

1) **Severely Limited Training Duration:** The most significant limitation was the premature conclusion of the PTDA fine-tuning process after only 5 epochs, despite a configuration targeting 20 epochs. The resulting high training loss (0.4919) indicates that the models were far from converged. This under-training profoundly affects the interpretation of PTDA's efficacy, as the observed negative or mixed results are likely attributable to the models being in an unstable, early phase of learning from the specialized synthetic data.
2) **Synthetic Speech Quality and Accent Modeling:** While Coqui TTS was used, the inherent limitations of current TTS technology in perfectly mimicking nuanced accents and the acoustic characteristics of child speech remain a challenge. The naturalness and precise phonetic realization of the generated French-accented Dutch speech could have influenced the fine-tuning process, potentially introducing artifacts or not fully capturing the targeted error patterns.
3) **Dataset Specificity and Size:** This research utilized the JASMIN-CGN corpus [10], which, while valuable for its inclusion of child and non-native speech, has specific recording conditions and a finite number of speakers in the target demographic groups (NL and FR children aged 7-11). The size of the PTDA dataset generated for fine-tuning was also constrained by the available source sentences from the Oscar corpus and the TTS generation process. Generalizability to other child speech datasets or broader real-world conditions might vary.
4) **Scope of Phoneme Analysis:** The phoneme error analysis, while detailed, primarily focused on PER. Other acoustic-phonetic features (e.g., duration, intonation, stress patterns) and prosodic elements also contribute significantly to accent perception and ASR recognition difficulty, which were not explicitly targeted by the PTDA approach.
5) **Computational Resource Constraints:** While the effective batch size was 16, the primary impact of resource constraints was the inability to complete the full planned training duration, which overshadowed other hyperparameter choices.

Future Work and Recommendations

Building upon the findings and limitations of this research, several avenues for future research and practical recommendations emerge to more robustly evaluate and potentially realize the benefits of PTDA:

- **Ensuring Sufficient Training and Convergence:** The foremost priority is to allow models to train for a sufficient duration to achieve convergence. This means completing the configured number of epochs (e.g., 20) or relying on early stopping criteria based on a representative validation set, rather than premature termination after only 5 epochs. Only by evaluating fully trained models can the true potential of PTDA be assessed. Monitoring validation loss and WER throughout training is important.
- **Refining Synthetic Data Generation and Exploring Advanced Synthesis Techniques:** While Coqui TTS was used in this research, the acoustic quality and naturalness of synthetic speech are of high importance. Future work should explore more advanced TTS systems and methodologies. For instance, fine-tuning or developing TTS models using toolkits like ESPnet, which offers extensive control over acoustic features and has shown strong performance in speech synthesis research, could be a promising avenue. This would allow for more nuanced control over the phoneme realizations and potentially higher fidelity accented speech generation, which is already being explored as a subsequent step. The quality of the voice cloning and its match to the target demographic (children) should also be carefully assessed.
- **Iterative Refinement and Broader Hyperparameter Exploration:** Beyond just achieving convergence, more general iteration on the fine-tuning process is needed once sufficient training is possible. This includes exploring a wider range of hyperparameters (e.g., different learning rates, more sophisticated schedulers, batch sizes if resources permit) and potentially different model adaptation techniques (e.g., LoRA, adapters) that might be more robust to limited data or more effective for bias mitigation tasks, especially when dealing with synthetic data.
- **Investigating Model-Specific Responses:** The differential response of Whisper Large-v3 and Large-v3-Turbo to the PTDA approach, even when under-trained, suggests model-specific sensitivities. Further investigation into why the Turbo model showed some absolute improvements while the standard Large-v3 degraded catastrophically could yield valuable insights for tailoring augmentation and fine-tuning strategies to specific model architectures, particularly when training to convergence.

Addressing these areas will be important for determining if PTDA can be a viable strategy for mitigating ASR bias in practical applications.

Revisiting Project Requirements

The project requirements, outlined in Table I, were evaluated based on the outcomes of the 5-epoch PTDA fine-tuning experiments:

1) **Req. 1: The phoneme-level bias between NL and FR children must be measurably reduced.** This requirement was **not met**. For both Whisper Large-v3 and Large-v3-Turbo models, the 5-epoch fine-tuning with PTDA generally led to an increase in performance disparities at both WER and PER levels, further disadvantaging the FR group. The under-trained state of the models likely prevented any potential bias reduction benefits and may have caused the models to learn spurious correlations that exacerbated bias.
2) **Req. 2: Overall WER for both speaker groups must not increase compared to baseline.** This requirement was **not met** for the Whisper Large-v3 model, where WERs increased dramatically for both groups. For the Whisper Large-v3-Turbo model, it was **partially met**, as mean WER decreased for both groups, but the median WER for the FR group increased.
3) **Req. 3: The solution must improve WER for both speaker groups, or at minimum not degrade performance for either group.** This requirement was **not met** for Whisper Large-v3 due to significant WER increases. For Whisper Large-v3-Turbo, it was **partially met**; mean WER improved for NL and marginally for FR, but median FR WER degraded.
4) **Req. 4: While reducing bias in targeted phonemes, the solution must not create new performance disparities in previously well-recognized phonemes.** This requirement was likely **not met**. For Whisper Large-v3, widespread PER increases suggest new disparities. For Whisper Large-v3-Turbo, shifts in Error_Diff for phonemes like /j/ and /a/ (from NL worse to FR worse) and degradation of /f/ indicate that new relative disparities or worsening of existing ones occurred.
5) **Req. 5: The methodology must be transferable to other ASR models beyond Whisper.** The direct evaluation of transferability was beyond the scope of these experiments. However, the catastrophic failure with Large-v3 and mixed results with Large-v3-Turbo, both stemming from under-training with Whisper models, suggest that significant challenges would exist in transferring this specific PTDA implementation without addressing the fundamental training issues. Thus, an evaluation is **inconclusive**.
6) **Req. 6: The synthetic data generation process must be transparent and explainable.** The methodology for PTDA, involving phoneme identification and targeted synthesis, was designed to be transparent and explainable. This aspect of the requirement concerning the process itself was **met**. However, the effectiveness of this transparent process in yielding positive ASR outcomes was severely limited by the under-training of the models.

Ultimately, the current findings underscore that the important factor of insufficient training duration prevented a fair evaluation of PTDA against most project requirements.

REFERENCES

[1] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards Inclusive Automatic Speech Recognition," *Computer Speech Language*, vol. 84, p. 101567, Mar. 2024. 10.1016/j.csl.2023.101567.

[2] M. Fuckner, S. Horsman, P. Wiggers, and I. Janssen, "Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch Speakers," in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, Romania: IEEE, Oct. 2023, pp. 146-151. 10.1109/SpeD59241.2023.10314895.

[3] M. K. Ngueajio and G. Washington, "Hey ASR System! Why Aren't You More Inclusive?: Automatic Speech Recognition Systems' Bias and Proposed Bias Mitigation Techniques. A Literature Review," in *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, vol. 13518, J. Y. C. Chen, G. Fragomeni, H. Degen, and S. Ntoa, Eds., Lecture Notes in Computer Science, vol. 13518, Cham: Springer Nature Switzerland, 2022, pp. 421-440. 10.1007/978-3-031-21707-$4_3$0.

[4] A. Renato, D. Luna, en S. Benítez, 'Development of an ASR System for Medical Conversations', in Studies in Health Technology and Informatics, J. Bichel-Findlay, P. Otero, P. Scott, en E. Huesing, Red., IOS Press, 2024. doi: 10.3233/SHTI231048.

[5] Centraal Bureau voor de Statistiek, "Taal en dialectgebruik; regio, 2019," CBS Open Data, 2020. [Online]. Available: https://opendata.cbs.nl//CBS/nl/dataset/84727NED/table. [Accessed: Mar. 14, 2025].

[6] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, "Wat zijn de erkende talen in Nederland?," Rijksoverheid, Jul. 2024. [Online]. Available: https://www.rijksoverheid.nl/onderwerpen/erkende-talen/vraag-en-antwoord/erkende-talen-nederland. [Accessed: Mar. 14, 2025].

[7] C.-T. Do, S. Imai, R. Doddipatla, and T. Hain, "Improving Accented Speech Recognition using Data Augmentation based on Unsupervised Text-to-Speech Synthesis," arXiv preprint *arXiv:2407.04047*, Jul. 2024. 10.48550/arXiv.2407.04047.

[8] Y. Zhang, A. Herygers, T. Patel, Z. Yue, and O. Scharenborg, "Exploring Data Augmentation in Bias Mitigation Against Non-Native-Accented Speech," arXiv preprint *arXiv:2312.15499*, Dec. 2023. 10.48550/arXiv.2312.15499.

[9] H. Su, H. Farn, F.-Y. Sun, S.-T. Chen, and H. Lee, "Task Arithmetic Can Mitigate Synthetic-to-Real Gap in Automatic Speech Recognition," arXiv preprint *arXiv:2406.02925*, Oct. 2024. 10.48550/arXiv.2406.02925.

[10] C. Cucchiarini, H. Van Hamme, O. van Herwijnen, and F. Smits, "JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children, and Non-Natives in the Human-Machine Interaction Modality," in *Proc. 5th International Conference on Language Resources and Evaluation (LREC'06)*, 2006.

[11] Hugging Face, "Transformers Documentation," *Hugging Face*, Accessed: May 19, 2025. [Online]. Available: https://huggingface.co/docs/transformers/en/index

[12] Hugging Face, "Fine-tuning the ASR model," *Hugging Face Audio Course*, Accessed: May 19, 2025. [Online]. Available: https://huggingface.co/learn/audio-course/chapter5/fine-tuning

[13] Ministerie van Justitie en Veiligheid, "Communication tools," Nationaal Coördinator Terrorismebestrijding en Veiligheid (NCTV), [Online]. Available: https://english.nctv.nl/topics/risk-and-crisis-communication/communication-tools. [Accessed: May 19, 2025].

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, and C. McLeavey, "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv preprint arXiv:2212.04356, Dec. 2022. 10.48550/arXiv.2212.04356.
[15] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," arXiv preprint *arXiv:2006.11477*, Jun. 2020. 10.48550/arXiv.2006.11477.
[16] D. de Vries, "The role of the Dutch government in the COVID-19 crisis: From crisis management to crisis governance," *Journal of Comparative Policy Analysis: Research and Practice*, vol. 24, no. 2, pp. 150–164, 2022. 10.1016/j.jcpa.2022.04.001.
[17] P. Ortiz Suarez, J. Abadji, R. Ismail, S. Takeshita, S. Nagel, and B. Sagot, "OSCAR 23.01: Open Super-large Crawled Aggregated coRpus," arXiv preprint arXiv:2301.00001, Jan. 2023. 10.48550/arXiv.2301.00001.
[18] Common Crawl Foundation, "Common Crawl: Open Repository of Web Crawl Data," 2025. [Online]. Available: https://commoncrawl.org/. [Accessed: May 19, 2025].
[19] eSpeak NG Developers, "eSpeak NG: Open Source Text-to-Speech Synthesizer," GitHub repository, [Online]. Available: https://github.com/espeak-ng/espeak-ng. [Accessed: May 19, 2025].
[20] Coqui.ai, "TTS: A deep learning toolkit for Text-to-Speech," GitHub repository, [Online]. Available: https://github.com/coqui-ai/TTS. [Accessed: May 19, 2025].
[21] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model," arXiv preprint *arXiv:2406.04904*, Jun. 2024. [Online]. Available: https://arxiv.org/abs/2406.04904.

APPENDIX

A. Whisper Large-v3 Model

TABLE X: Full Per-Phoneme Statistics for Whisper Large-v3 Model. Phoneme symbols are cleaned (braces removed). Error Diff = FR_PER - NL_PER.

Phoneme	NL PER (%)	FR PER (%)	Error Diff (%)	NL Count	FR Count	NL Freq (%)	FR Freq (%)	Avg Freq (%)	Importance Score
ₙ	14.67	11.18	−3.50	19 374	5171	9.2627	9.2143	9.2385	−106.28
@	14.67	11.65	−3.01	21 714	5758	10.3814	10.2603	10.3209	−96.85
r	12.82	10.56	−2.25	12 891	3436	6.1632	6.1227	6.1429	−55.78
O	14.39	10.99	−3.40	4713	1174	2.2533	2.0920	2.1726	−50.08
j	18.09	13.88	−4.20	2449	677	1.1709	1.2064	1.1886	−45.84
A	12.64	10.42	−2.22	8289	2207	3.9630	3.9327	3.9478	−44.15
d	13.57	11.43	−2.14	9078	2300	4.3402	4.0984	4.2193	−43.89
f	18.49	14.33	−4.16	2147	642	1.0265	1.1440	1.0852	−43.34
t	12.44	11.00	−1.44	18 659	4893	8.9208	8.7190	8.8199	−42.88
h	15.68	13.08	−2.60	4394	1185	2.1008	2.1116	2.1062	−37.74
a	14.74	12.86	−1.88	6655	1851	3.1817	3.2983	3.2400	−33.89
Z	52.17	20.00	−32.17	23	5	0.0110	0.0089	0.0100	−32.10
E+	13.38	15.71	2.33	3708	993	1.7728	1.7695	1.7711	31.05
z	14.74	12.83	−1.90	4927	1395	2.3556	2.4858	2.4207	−29.62
Y+	13.50	16.85	3.35	1067	279	0.5101	0.4972	0.5036	23.77
g	0.00	25.00	25.00	19	4	0.0091	0.0071	0.0081	22.51
w	11.75	10.14	−1.61	3634	986	1.7374	1.7570	1.7472	−21.26
G	13.56	11.83	−1.73	3172	837	1.5165	1.4915	1.5040	−21.19
m	11.57	10.32	−1.25	5956	1599	2.8476	2.8493	2.8484	−21.08
v	13.52	12.04	−1.48	3801	1013	1.8173	1.8051	1.8112	−19.91
s	13.78	12.84	−0.93	8994	2297	4.3000	4.0931	4.1966	−19.11
u	12.48	10.66	−1.81	2020	544	0.9658	0.9694	0.9676	−17.84
x	13.62	12.34	−1.28	3759	1029	1.7972	1.8336	1.8154	−17.23
b	13.47	12.16	−1.31	3295	855	1.5753	1.5235	1.5494	−16.32
A+	14.67	17.00	2.33	825	200	0.3944	0.3564	0.3754	14.30
y	12.45	10.06	−2.39	707	179	0.3380	0.3190	0.3285	−13.70
2	16.30	19.20	2.90	460	125	0.2199	0.2227	0.2213	13.62
N	11.24	13.04	1.81	1059	276	0.5063	0.4918	0.4991	12.76
i	15.09	14.29	−0.80	3559	959	1.7016	1.7089	1.7052	−10.48
o	12.84	12.10	−0.73	4385	1107	2.0965	1.9726	2.0345	−10.48
E	13.13	12.50	−0.63	5554	1520	2.6554	2.7085	2.6819	−10.25
l	11.95	11.59	−0.36	8437	2304	4.0337	4.1056	4.0696	−7.24
I	13.38	12.96	−0.42	5381	1543	2.5726	2.7495	2.6611	−6.83
S	38.18	40.74	2.56	110	27	0.0526	0.0481	0.0504	5.74
p	12.52	12.09	−0.43	3658	984	1.7489	1.7534	1.7512	−5.65
k	11.38	11.15	−0.23	6602	1812	3.1564	3.2289	3.1926	−4.06
Y	16.15	16.37	0.22	1542	397	0.7372	0.7074	0.7223	1.91
e	14.15	14.07	−0.08	3994	1109	1.9095	1.9762	1.9428	−1.11

B. Whisper Large-v3-Turbo Model

TABLE XI: Full Per-Phoneme Statistics for Whisper Large-v3-Turbo Model. Phoneme symbols are cleaned. Error Diff = FR_PER - NL_PER.

Phoneme	NL PER (%)	FR PER (%)	Error Diff (%)	NL Count	FR Count	NL Freq (%)	FR Freq (%)	Avg Freq (%)	Importance Score
j	33.07	26.14	−6.93	2449	677	1.1839	1.2287	1.2063	−76.11
d	28.46	25.78	−2.68	9078	2300	4.3885	4.1742	4.2813	−55.49
f	32.84	27.57	−5.27	2147	642	1.0379	1.1652	1.1015	−55.27
a	30.22	27.23	−2.99	6655	1851	3.2171	3.3593	3.2882	−54.21
N	28.05	35.51	7.46	1059	276	0.5119	0.5009	0.5064	53.10
n	28.35	27.07	−1.27	19 374	5171	9.3657	9.3848	9.3752	−38.98
v	29.15	26.46	−2.69	3801	1013	1.8375	1.8385	1.8380	−36.53
r	28.83	27.47	−1.36	12 891	3436	6.2317	6.2359	6.2338	−33.96
m	27.80	25.89	−1.91	5956	1599	2.8792	2.9020	2.8906	−32.52
o	30.44	28.18	−2.26	4385	1107	2.1198	2.0091	2.0644	−32.48
y	27.86	32.96	5.10	707	179	0.3418	0.3249	0.3333	29.43
b	30.05	32.16	2.12	3295	855	1.5929	1.5517	1.5723	26.56
A	29.28	28.05	−1.23	8289	2207	4.0070	4.0054	4.0062	−24.67
t	28.32	27.61	−0.71	18 659	4893	9.0201	8.8802	8.9501	−21.34
Z	39.13	20.00	−19.13	23	5	0.0111	0.0091	0.0101	−19.22
s	29.38	30.26	0.88	8994	2297	4.3478	4.1688	4.2583	18.19
E+	29.48	30.82	1.34	3708	993	1.7925	1.8022	1.7973	17.95
i	31.53	30.24	−1.29	3559	959	1.7205	1.7405	1.7305	−16.92
z	30.02	29.03	−0.99	4927	1395	2.3818	2.5318	2.4568	−15.45
k	28.19	28.92	0.73	6602	1812	3.1915	3.2886	3.2400	13.14
p	30.34	31.30	0.96	3658	984	1.7683	1.7858	1.7771	12.75
x	29.53	30.42	0.89	3759	1029	1.8172	1.8675	1.8423	12.06
l	28.42	28.95	0.53	8437	2304	4.0786	4.1815	4.1300	10.71
Y+	29.80	31.18	1.38	1067	279	0.5158	0.5064	0.5111	9.86
e	30.10	29.49	−0.61	3994	1109	1.9308	2.0127	1.9717	−8.55
u	29.65	30.33	0.68	2020	544	0.9765	0.9873	0.9819	6.71
@	27.93	27.74	−0.20	21 714	5758	10.4969	10.4501	10.4735	−6.34
g	31.58	25.00	−6.58	19	4	0.0092	0.0073	0.0082	−5.97
G	29.04	29.51	0.47	3172	837	1.5334	1.5191	1.5262	5.87
2	31.74	32.80	1.06	460	125	0.2224	0.2269	0.2246	5.03
h	30.00	30.21	0.22	4394	1185	2.1241	2.1506	2.1374	3.15
O	29.03	29.22	0.19	4713	1174	2.2783	2.1307	2.2045	2.82
S	47.27	48.15	0.88	110	27	0.0532	0.0490	0.0511	1.98
E	29.51	29.41	−0.10	5554	1520	2.6849	2.7586	2.7218	−1.69
I	28.10	28.19	0.09	5381	1543	2.6013	2.8004	2.7008	1.53
w	29.61	29.72	0.11	3634	986	1.7567	1.7895	1.7731	1.42
A+	33.33	33.50	0.17	825	200	0.3988	0.3630	0.3809	1.03
Y	30.67	30.73	0.06	1542	397	0.7454	0.7205	0.7330	0.48

1) Whisper Large-v3 Model (Fine-tuned with PTDA after 5 Epochs):

TABLE XII: Full Per-Phoneme Statistics for Whisper Large-v3 Model Fine-tuned with PTDA (5 Epochs). Phoneme symbols are shown as in data. Error Diff = FR_PER - NL_PER.

Phoneme	NL PER (%)	FR PER (%)	Error Diff (%)	NL Count	FR Count	NL Freq (%)	FR Freq (%)	Avg Freq (%)	Importance Score
s	54.55	288.89	234.34	110	27	0.0505	0.0460	0.0482	514.57
I	25.83	46.66	20.83	5381	1543	2.4692	2.6261	2.5476	332.49
t	23.89	34.40	10.50	18 659	4893	8.5622	8.3275	8.4449	305.25
k	26.83	43.76	16.94	6602	1812	3.0295	3.0839	3.0567	296.14
l	24.81	38.67	13.86	8437	2304	3.8716	3.9212	3.8964	273.67
A+	24.97	69.50	44.53	825	200	0.3786	0.3404	0.3595	266.99
A	26.76	39.96	13.21	8289	2207	3.8036	3.7561	3.7799	256.74
n	25.00	33.42	8.41	19 374	5171	8.8903	8.8007	8.8455	250.26
@	25.30	33.22	7.92	21 714	5758	9.9641	9.7997	9.8819	249.02
s	27.44	39.83	12.39	8994	2297	4.1271	3.9093	4.0182	248.45
O	24.49	41.65	17.17	4713	1174	2.1627	1.9981	2.0804	247.61
a	26.97	40.73	13.76	6655	1851	3.0538	3.1503	3.1020	242.39
E+	27.94	45.12	17.18	3708	993	1.7015	1.6900	1.6958	223.67
r	26.38	35.45	9.07	12 891	3436	5.9154	5.8478	5.8816	219.86
h	27.54	42.87	15.33	4394	1185	2.0163	2.0168	2.0165	217.72
m	24.78	37.90	13.12	5956	1599	2.7331	2.7214	2.7272	216.62

Continued on next page

TABLE XII: Full Per-Phoneme Statistics for Whisper Large-v3 Model Fine-tuned with PTDA (5 Epochs) (Continued)

Phoneme	NL PER (%)	FR PER (%)	Error Diff (%)	NL Count	FR Count	NL Freq (%)	FR Freq (%)	Avg Freq (%)	Importance Score
i	27.40	44.11	16.71	3559	959	1.6331	1.6321	1.6326	213.55
v	27.41	43.44	16.02	3801	1013	1.7442	1.7240	1.7341	210.98
N	25.50	55.43	29.94	1059	276	0.4860	0.4697	0.4778	206.96
b	25.58	42.46	16.87	3295	855	1.5120	1.4551	1.4836	205.50
d	24.03	34.17	10.15	9078	2300	4.1657	3.9144	4.0401	203.99
E	28.47	39.93	11.47	5554	1520	2.5486	2.5869	2.5678	183.77
p	26.00	39.74	13.74	3658	984	1.6786	1.6747	1.6766	177.89
y	23.90	53.07	29.17	707	179	0.3244	0.3046	0.3145	163.59
j	32.42	47.56	15.14	2449	677	1.1238	1.1522	1.1380	161.52
G	24.75	38.11	13.36	3172	837	1.4556	1.4245	1.4400	160.38
x	28.97	40.82	11.85	3759	1029	1.7249	1.7513	1.7381	156.17
Y	28.73	46.35	17.62	1542	397	0.7076	0.6757	0.6916	146.52
Y+	24.46	45.52	21.06	1067	279	0.4896	0.4748	0.4822	146.24
f	33.21	47.35	14.14	2147	642	0.9852	1.0926	1.0389	144.16
w	28.67	39.35	10.68	3634	986	1.6676	1.6781	1.6728	138.10
2	30.87	60.00	29.13	460	125	0.2111	0.2127	0.2119	134.10
e	27.87	37.33	9.46	3994	1109	1.8328	1.8874	1.8601	129.08
z	28.37	36.27	7.90	4927	1395	2.2609	2.3742	2.3175	120.24
o	32.06	40.29	8.23	4385	1107	2.0122	1.8840	1.9481	114.80
u	28.91	39.15	10.24	2020	544	0.9269	0.9258	0.9264	98.59
g	26.32	50.00	23.68	19	4	0.0087	0.0068	0.0078	20.87
Z	69.57	80.00	10.43	23	5	0.0106	0.0085	0.0095	10.19

2) Whisper Large-v3-Turbo Model (Fine-tuned with PTDA after 5 Epochs):

TABLE XIII: Full Per-Phoneme Statistics for Whisper Large-v3-Turbo Model Fine-tuned with PTDA (5 Epochs). Phoneme symbols are shown as in data. Error Diff = FR_PER - NL_PER.

Phoneme	NL PER (%)	FR PER (%)	Error Diff (%)	NL Count	FR Count	NL Freq (%)	FR Freq (%)	Avg Freq (%)	Importance Score
Y+	22.77	41.94	19.16	1067	279	0.5087	0.5025	0.5056	136.25
I	19.66	26.12	6.46	5381	1543	2.5654	2.7791	2.6723	105.54
x	19.21	26.43	7.23	3759	1029	1.7921	1.8534	1.8227	97.56
E	18.19	23.95	5.76	5554	1520	2.6479	2.7377	2.6928	94.56
G	19.07	26.76	7.69	3172	837	1.5123	1.5075	1.5100	94.48
l	18.08	22.31	4.23	8437	2304	4.0224	4.1498	4.0861	85.58
E+	20.28	26.49	6.20	3708	993	1.7678	1.7885	1.7782	82.74
O	18.67	24.02	5.35	4713	1174	2.2470	2.1145	2.1807	78.99
s	20.93	24.64	3.72	8994	2297	4.2880	4.1372	4.2126	76.26
t	18.17	20.72	2.56	18 659	4893	8.8958	8.8129	8.8544	76.04
e	18.68	24.08	5.40	3994	1109	1.9042	1.9974	1.9508	75.39
A+	45.09	33.00	−12.09	825	200	0.3933	0.3602	0.3768	−74.22
m	16.19	20.14	3.95	5956	1599	2.8396	2.8800	2.8598	66.84
N	18.89	28.26	9.38	1059	276	0.5049	0.4971	0.5010	66.36
w	17.97	22.92	4.95	3634	986	1.7325	1.7759	1.7542	65.58
p	17.93	22.76	4.83	3658	984	1.7440	1.7723	1.7581	64.06
k	18.48	21.91	3.43	6602	1812	3.1476	3.2636	3.2056	61.42
g	21.05	80.00	58.95	19	5	0.0091	0.0090	0.0090	56.02
A	17.81	20.62	2.81	8289	2207	3.9518	3.9751	3.9635	55.93
f	36.98	31.93	−5.05	2147	642	1.0236	1.1563	1.0900	−52.73
b	21.37	25.15	3.78	3295	855	1.5709	1.5400	1.5554	47.15
a	19.13	21.66	2.54	6655	1851	3.1728	3.3339	3.2533	45.73
i	22.00	25.44	3.44	3559	959	1.6968	1.7273	1.7120	45.04
j	32.95	36.04	3.09	2449	677	1.1676	1.2194	1.1935	33.75
u	20.05	23.35	3.30	2020	544	0.9631	0.9798	0.9714	32.49
@	19.76	20.74	0.97	21 714	5758	10.3523	10.3709	10.3616	31.38
h	20.46	22.53	2.07	4394	1185	2.0949	2.1343	2.1146	30.13
S	47.27	59.26	11.99	110	27	0.0524	0.0486	0.0505	26.95
v	19.86	21.32	1.46	3801	1013	1.8122	1.8245	1.8183	19.68
z	19.59	20.65	1.06	4927	1395	2.3490	2.5126	2.4308	16.51
r	21.66	21.07	−0.59	12 891	3436	6.1459	6.1886	6.1673	−14.59
n	18.50	18.86	0.35	19 374	5171	9.2367	9.3136	9.2752	10.69
Z	30.43	40.00	9.57	23	5	0.0110	0.0090	0.0100	9.56

Continued on next page

TABLE XIII: Full Per-Phoneme Statistics for Whisper Large-v3-Turbo Model Fine-tuned with PTDA (5 Epochs) (Continued)

Phoneme	NL PER (%)	FR PER (%)	Error Diff (%)	NL Count	FR Count	NL Freq (%)	FR Freq (%)	Avg Freq (%)	Importance Score
Y	35.41	36.52	1.12	1542	397	0.7352	0.7150	0.7251	9.50
o	21.14	21.68	0.54	4385	1107	2.0906	1.9938	2.0422	7.72
d	18.81	18.96	0.14	9078	2300	4.3280	4.1426	4.2353	2.92
y	22.49	22.91	0.42	707	179	0.3371	0.3224	0.3297	2.39
2	35.43	35.20	−0.23	460	125	0.2193	0.2251	0.2222	−1.11