Universiteit Leiden

Master Computer Science

From Speaker-Independent to Speaker-Specific: Two-Stage Whisper Adaptation for Aphasia Speech Recognition

Name:	Shan Jiang
Student ID:	s3824322
Date:	28/01/26
Specialisation:	Artificial Intelligence
1st supervisor:	Dr. Qinyu Chen
2nd supervisor:	Dr. Yingqiang Gao

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Automatic speech recognition (ASR) is widely used in many domains, yet it often performs poorly on impaired speech, limiting voice-based assistive communication. This thesis investigates three questions: (1) whether Speaker-Independent Fine-tuning (SI-FT) can reduce word error rate (WER), (2) whether Speaker-Specific Fine-tuning (SS-FT) initialized from a speaker-independent model outperforms direct personalization from the Whisper baseline, and (3) whether lower WER is associated with improved downstream recovery of the speaker's intended meaning by large language models (LLMs).

To answer these questions, a two-stage adaptation framework is developed on top of Whisper. The system performs SI-FT followed by SS-FT, and compares four ASR conditions: Whisper baseline, SI-FT, SI-FT→SS-FT, and Whisper→SS-FT. Experiments use a speaker-disjoint protocol on AphasiaBank conversational speech. In addition, TED-LIUM v3 and FLEURS are used as out-of-domain checks to quantify typical-speech degradation after adaptation. For downstream evaluation, an exploratory LLM-based semantic similarity scoring protocol is applied to compare ASR hypotheses with reference transcripts as a proxy measure of semantic fidelity.

Results show that SI-FT substantially reduces WER on impaired speech, demonstrating strong cold-start gains, while out-of-domain effects on typical speech remain limited in this setting. For personalization, the two-stage strategy SI-FT→SS-FT yields lower WER than direct Whisper→SS-FT for most target speakers, indicating that SI-FT provides a better prior for speaker-specific adaptation. Finally, although overall similarity means differ only slightly across ASR conditions, per-speaker analyses suggest a negative association between WER and semantic similarity scores, pointing to a link between transcription accuracy and downstream recovery of intended meaning.

Contents

Abstract **ii**

1 Introduction **1**

1.1 Motivation 1

1.2 Problem Statement 2

1.3 Proposed Approach 2

1.4 Thesis Structure 3

2 Related Work **4**

2.1 Aphasia-related speech and the limitations of general ASR 4

2.2 Whisper: Robust Speech Recognition via Large-Scale Weak Supervision 5

2.3 Adaptation routes for disordered speech 5

2.4 Read speech versus conversational speech 5

2.5 Adaptation and out-of-domain degradation control 6

2.6 Summary 6

3 Methodology **7**

3.1 Task Definition 7

3.2 System Overview 7

3.3 Experimental Conditions 9

3.4 Method Details 10

3.4.1 Input 10

3.4.2 Speaker-Independent Fine-tuning(SI-FT;B2) 11

3.4.3 Speaker-Specific Fine-tuning(SS-FT;B3 and B4) 11

3.4.4 LLM-based semantic similarity scoring 12

3.4.5 Evaluation 12

4 Data **14**
4.1 Datasets . . . 14
4.1.1 AphasiaBank . . . 14
4.1.2 TED-LIUM v3 . . . 15
4.1.3 FLEURS . . . 15
4.2 Preprocessing . . . 15
4.2.1 Audio preprocessing . . . 16
4.2.2 Transcript preprocessing . . . 16
4.3 Data splits . . . 17
4.4 Leakage Prevention . . . 20

5 Experiments **21**
5.1 Experimental Setup . . . 21
5.1.1 Compute environment. . . . 21
5.1.2 Training Setup . . . 21
5.1.3 Decoding configuration . . . 22
5.2 Baselines . . . 23
5.3 Experimental protocol . . . 23
5.4 Metrics . . . 24
5.4.1 Word Error Rate (WER) . . . 24
5.4.2 LLM-based Similarity . . . 24
5.5 RQs experiments . . . 25
5.5.1 RQ1: SI-FT vs Whisper baseline . . . 26
5.5.2 RQ2: SI-FT→SS-FT vs Whisper→SS-FT . . . 26
5.5.3 RQ3: LLM-based similarity scoring . . . 26

6 Results and Discussion **27**
6.1 RQ1: SI-FT vs Whisper baseline . . . 27
6.2 RQ2: SI-FT→SS-FT vs Whisper→SS-FT . . . 29
6.3 RQ3: LLM-based similarity scoring . . . 32

7 Conclusions **36**
7.1 Findings . . . 36
7.2 Limitations . . . 37
7.3 Future work . . . 38

A SS-FT Dataset Details **40**

B Hyperparameters **42**

C Text normalization for WER **43**

C.1 LLM-Based Similarity Evaluation Prompt 44

C.1.1 Prompt Template 44

C.1.2 Prompt Construction 44

C.1.3 Definition of `intro` 44

D Per-speaker results for RQ2 **45**

Bibliography **47**

Chapter 1

Introduction

1.1 Motivation

More than 170 million people worldwide live with some form of speech impairment [11]. Disorders such as aphasia, dysarthria, and related speech and language disorders can affect both the acoustic speech signal and language production. Under these conditions, general-purpose automatic speech recognition (ASR) systems often exhibit much higher error rates, producing wrong words, omissions, or hallucinated content. These errors can in turn trigger semantic misunderstandings and degrade subsequent human-computer interaction.

In parallel, large language models (LLMs) have become widely used for conversational interaction and language understanding, and are increasingly considered for assistive communication and clinical text processing. However, speech-based interaction between impaired speakers and LLMs remains difficult when the upstream ASR is inaccurate. This motivates a practical question: whether improvements in ASR quality can consistently translate into better downstream semantic understanding and interaction quality. Moreover, the inference cost and deployment constraints of LLMs are non-negligible in practical applications. If improved ASR can reduce reliance on very large LLMs, smaller models may achieve comparable semantic understanding, which can enable on-device deployment, improve privacy, and support offline use.

1.2 Problem Statement

A common direction in impaired-speech ASR is speaker-specific personal adaptation(PA), where a model is adapted to a single target speaker using a small amount of speaker-specific data. This approach can yield substantial gains in many settings, but it faces two major limitations. First, cold start is difficult to avoid. New users typically lack sufficient personal data, and collecting and transcribing speaker-specific training data can be costly. Second, personal data collection often relies on reading scripted sentences. This can introduce a mismatch to real conversational communication, which may reduce effectiveness and limit generalization in interactive scenarios.

Based on these observations, this thesis investigates the following research questions:

- **RQ1:** Can Speaker-Independent Fine-tuning (SI-FT) significantly reduce WER on impaired speech?
- **RQ2:** Does performing speaker-specific fine-tuning after SI-FT yield better personalization results than personal adaptation starting directly from the Whisper [10] baseline?
- **RQ3:** Do ASR improvements lead to better semantic recovery of the speaker's intended meaning, enabling downstream LLMs to infer the speaker's intended meaning more accurately? Can stronger ASR reduce the gap between smaller and larger LLMs?

1.3 Proposed Approach

To investigate these questions, this thesis proposes and evaluates a two-stage adaptation framework based on Whisper.

Stage 1 performs Speaker-Independent Fine-tuning (SI-FT) using impaired-speech data from multiple speakers. The goal is to learn cross-speaker pronunciation and acoustic variability and to produce a stronger cold-start ASR model for the target population. Stage 2 performs Speaker-Specific Fine-tuning (SS-FT) on speaker-specific data. For each target speaker, personalization is performed and two initialization strategies are compared: initializing from the SI-FT model versus initializing from the Whisper baseline.

A strict speaker-disjoint data split is used to construct four experimental conditions. Building on ASR evaluation, transcription outputs from each condition are

used as inputs to LLMs of different sizes. An LLM-based semantic similarity scoring protocol is applied against reference text to approximate downstream interpretability, providing an exploratory signal of whether ASR improvements can benefit semantic understanding.

1.4 Thesis Structure

This thesis is organized as follows. Chapter 2 reviews related work on foundation ASR models, impaired-speech adaptation and personalization, conversational evaluation settings, and strategies for mitigating out-of-domain degradation. Chapter 3 introduces the system overview, experimental conditions, and method details. Chapter 4 presents the datasets used in this thesis, including audio preprocessing, transcript preprocessing, and data splits. Chapter 5 details the experimental setup, conditions, and metrics. Chapter 6 reports and analyzes the experimental results in relation to the research questions. Finally, Chapter 7 concludes the thesis and discusses limitations and future directions.

Chapter 2

Related Work

2.1 Aphasia-related speech and the limitations of general ASR

Aphasia is typically associated with brain injury and affects language production and comprehension. Spoken output often exhibits word-finding difficulty, omission of content words, simplified syntax, increased pauses, repetition, and self-repair. Reduced intelligibility in aphasia-related speech commonly reflects two co-existing sources of difficulty. One source is a motor speech impairment, such as dysarthria or apraxia of speech, which leads to imprecise articulation and degraded pronunciation, producing substantial acoustic and phonological deviation from typical speech. The other source is the language impairment itself, which can manifest as lexical retrieval failures, semantic substitutions, and grammatical abnormalities, making the intended content harder to infer. In real communication settings, these difficulties are compounded by the natural properties of conversational speech, including disfluencies, corrections, repetition, and fragmented responses, which further increases the difficulty of ASR.

General ASR systems are typically trained on typical speech distributions. Their acoustic representations and language priors may not cover the high variability of pathological speech, leading to substantially higher error rates for aphasia and related disordered speech. This gap is further amplified when the evaluation target shifts from read speech to spontaneous conversation, where changes in linguistic structure and interaction dynamics make recognition more challenging.

2.2 Whisper: Robust Speech Recognition via Large-Scale Weak Supervision

Whisper is one of the most widely used ASR models in recent work. It is trained at scale on approximately 680,000 hours of multilingual and multitask supervision and demonstrates strong zero-shot robustness across a range of benchmarks [10]. Due to its unified foundation model and mature engineering ecosystem, Whisper is commonly used as a starting point for pathological speech ASR and can be adapted to specific populations or scenarios via fine-tuning. This thesis adopts Whisper as the base ASR model and focuses on adaptation strategies and evaluation in aphasia-related speech.

2.3 Adaptation routes for disordered speech

Prior work shows that targeted fine-tuning of Whisper can substantially improve WER on post-stroke and aphasia-related speech, with evidence of partial transfer across datasets [12]. These results provide evidence that Whisper's limitations on pathological speech can be mitigated through speaker-independent adaptation, which can serve as a stronger cold-start model for the target population.

Beyond domain mismatch at the population level, disordered speech also exhibits large cross-speaker variability. A complementary line of research therefore emphasizes speaker-specific personalization. Jordan and colleagues report that personalized models can outperform human listeners on short-phrase recognition, supporting the effectiveness of personalization for disordered speech [4]. Such findings indicate that speaker-specific fine-tuning can be highly effective when speaker-specific data are available. However, personalization also faces practical constraints. It often requires a non-trivial amount of high-quality user data, and the data collection process commonly relies on reading fixed phrases, which may not match conversational distributions and may exacerbate cold-start burden.

2.4 Read speech versus conversational speech

In disordered-speech ASR, the type of training and evaluation speech has a strong impact on real-world performance. Prior work shows that models trained or adapted primarily on read speech often perform substantially worse on conversational speech [13]. Conversational speech more frequently includes incomplete utterances, repetition, self-

repair, stronger contextual dependence, and more complex pause patterns.

Motivated by this, this thesis extracts utterance-level samples from AphasiaBank conversational recordings for training and evaluation. A topic-based strategy is used in constructing the SS-FT test set to better reflect realistic communication as topic-oriented responses, enabling a more meaningful assessment of the potential benefits of adaptation in practical use.

2.5 Adaptation and out-of-domain degradation control

Adaptation to impaired speech can introduce degradation on typical speech, for example due to domain overfitting or catastrophic forgetting. Bao and colleagues proposes hybrid fine-tuning strategies that mix pathological and typical speech data to improve aphasic speech recognition while maintaining performance on typical speech [1]. In practical assistive settings, maintaining typical speech capability is also relevant, for example when a patient's speech changes over time during recovery.

2.6 Summary

In summary, large-scale weakly supervised pretraining provides a strong foundation for robust ASR, but performance on aphasia-related speech remains substantially worse than on typical speech. Existing studies support the feasibility of adapting foundation ASR models to aphasic speech and validate the effectiveness of personalization, but cold-start burden and data mismatch remain important concerns. Conversational-speech studies further indicate that conclusions drawn from read speech may not transfer to real interaction settings. In addition, adaptation methods should mitigate performance degradation when applied to out-of-domain data.

Based on these observations, this thesis proposes and evaluates a two-stage adaptation framework under a strict speaker-disjoint protocol. Four training conditions are compared, with additional out-of-domain evaluations. An exploratory LLM-based semantic similarity analysis further tests whether improved ASR accuracy helps downstream models recover the speaker's intended meaning.

Chapter 3

Methodology

3.1 Task Definition

This thesis studies automatic speech recognition (ASR) for impaired speech in aphasia and related pronunciation disorders. Four training strategies are compared in terms of transcription accuracy. The baseline definitions are provided in Section 3.3.

The task is defined at the utterance level. Given an audio segment x, an ASR system produces a text transcription $\hat{y}$. A human transcript y is provided as the reference. Across all strategies, the goal is to generate a more accurate transcription $\hat{y}$ for impaired speech.

Performance is mainly evaluated using word error rate (WER), which measures the edit distance between the hypothesis $\hat{y}$ and the reference y. In addition to ASR accuracy, this thesis includes an exploratory downstream analysis of semantic fidelity. The ASR output $\hat{y}$ is used as input to large language models (LLMs) of different sizes, which provide a semantic similarity or comprehensibility score as a proxy measure of semantic similarity beyond literal matching.

3.2 System Overview

Speech intelligibility degradation in people with aphasia often arises from two co-existing sources. First, motor speech disorders such as dysarthria and apraxia of speech can lead to impaired articulation, causing speech to become difficult to understand at the acoustic and phonetic level. Second, aphasia-related language impairments can lead to word-finding difficulties, semantic substitutions, and grammatical

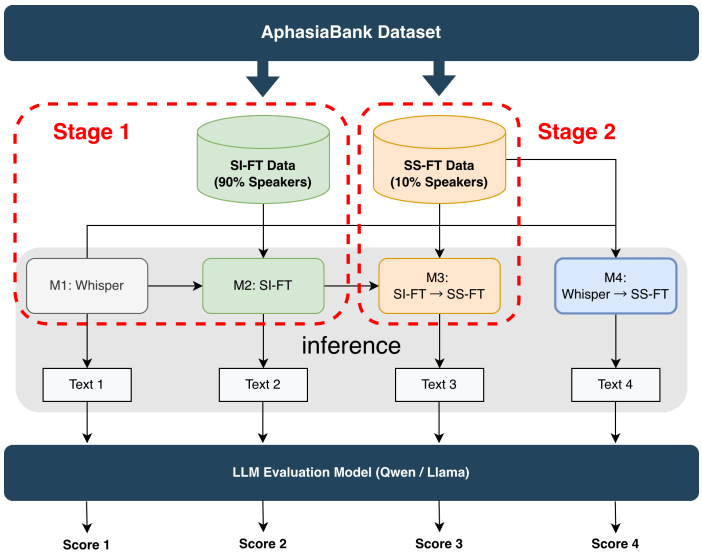


Figure 3.1: Overview of the two-stage ASR adaptation and downstream semantic evaluation pipeline. AphasiaBank speakers are split in a speaker-disjoint manner into Stage 1 SI-FT data and Stage 2 SS-FT data. Four ASR model conditions are compared: M1 (vanilla Whisper), M2 (SI-FT), M3 (SI-FT→SS-FT), and M4 (Whisper→SS-FT). Each model produces a transcript, which is then scored by LLM-based semantic similarity evaluation using Qwen and LLaMA models.

abnormalities, making the content harder to interpret. This thesis focuses on the first aspect. By selecting utterances with pronunciation errors only and adopting a two-stage adaptation framework, the study aims to address the cold-start challenge when speaker-specific data are limited and to further adapt to speaker variability and potential changes over time, improving ASR accuracy. In addition, this thesis explores whether higher transcription accuracy also improves semantic fidelity. To this end, LLMs of different sizes are used for an exploratory comprehensibility evaluation as an initial validation of potential downstream use.

Figure 3.1 illustrates the overall pipeline studied in this thesis. The system follows a two-stage adaptation framework built on Whisper, consisting of Speaker-Independent Fine-tuning (SI-FT) and Speaker-Specific Fine-tuning (SS-FT). Whisper [10] is a transformer-based automatic speech recognition model trained on large-scale multilingual and multitask speech data, and it provides strong zero-shot transcription performance that can be further improved through fine-tuning for specific domains such as impaired speech.

Prior work on impaired-speech ASR commonly starts from models trained on typi-

cal speech and then applies speaker-specific adaptation. This setting faces a cold-start problem. Building a usable personalized model may require substantial high-quality recordings. For instance, Project Relate [3] indicates that users may need to record 500 phrases to build a personalized model. This motivates Stage 1, where SI-FT adapts Whisper using impaired-speech data from multiple speakers to learn cross-speaker pronunciation patterns and provide a stronger starting point. Building on SI-FT, Stage 2 performs SS-FT to further adapt the model to a single target speaker, since individual speakers can exhibit distinct and time-varying characteristics.

The four experimental conditions are defined in Section 3.3. These conditions isolate the contribution of speaker-independent adaptation and test whether SI-FT provides a better initialization for speaker-specific fine-tuning.

An exploratory LLM-based similarity scoring is used to probe semantic fidelity.

The detailed dataset splits, training configurations, and evaluation protocols are described in Chapters 4 and 5.

3.3 Experimental Conditions

To isolate the effect of each adaptation stage, this thesis compares four ASR conditions based on Whisper. All conditions share the same data preprocessing, decoding configuration, and evaluation protocol; they differ only in the fine-tuning strategy and the data used for adaptation.

- **B1: Whisper (baseline).** The original Whisper model is used without any fine-tuning on impaired speech. This condition serves as the zero-shot baseline.
- **B2: SI-FT.** Whisper is adapted using Speaker-Independent Fine-tuning (SI-FT) on impaired-speech data from multiple speakers, resulting in a general model for the target population.
- **B3: SI-FT→SS-FT (two-stage).** Speaker-Specific Fine-tuning (SS-FT) is performed for each target speaker, initialized from the SI-FT model in B2.
- **B4: Whisper→SS-FT (direct adaptation).** Speaker-Specific Fine-tuning (SS-FT) is performed for each target speaker, initialized directly from the Whisper baseline in B1.

B1 and B2 are used to evaluate the benefit of speaker-independent adaptation (RQ1). B3 and B4 are used to evaluate whether SI-FT provides a better initialization for speaker-specific fine-tuning (RQ2). For the exploratory downstream analysis

(RQ3), transcripts produced under B1–B4 are compared using LLM-based semantic similarity scoring.

3.4 Method Details

This section describes the training and evaluation pipeline for the two-stage adaptation framework, including Speaker-Independent Fine-tuning (SI-FT), Speaker-Specific Fine-tuning (SS-FT), and the basic setup for LLM-based similarity scoring. Both stages are built on Whisper and are implemented as supervised sequence-to-sequence fine-tuning, where the model learns to map utterance-level audio features to text by minimizing the generation loss on paired audio and transcripts. Dataset construction and preprocessing are described in Chapter 4, while hyperparameter values and per-experiment configurations are summarized in Chapter 5.

Base model choice. This work fine-tunes `openai/whisper-small.en` as the base model. This choice is motivated by two considerations. First, the target language in this thesis is English, so an English-only Whisper variant is appropriate. Second, the *small* model offers a practical balance between recognition accuracy and inference cost. Larger Whisper variants generally improve WER, but the marginal gains beyond small diminish relative to the substantial increase in parameters and runtime. Table 3.1 reports representative WER numbers from [10], motivating our selection of `small.en`.

Training and decoding implementation Training is implemented with Hugging Face Transformers using `Seq2SeqTrainer`[14]. Evaluation is performed with the official OpenAI Whisper decoder after checkpoint conversion (Section 3.4.5).

3.4.1 Input

All audio is resampled to 16 kHz and converted to Whisper log-Mel spectrogram features using `WhisperFeatureExtractor`. Reference transcripts are tokenized with `WhisperTokenizer` to obtain label token IDs. Variable-length label sequences are padded to a common length, and padding positions are set to -100 so they do not contribute to the training loss. If a label sequence starts with a beginning-of-sentence (BOS) token, it is removed to avoid duplicating this token that is automatically added by the decoder. For efficiency, input features and label IDs are precomputed and cached in Arrow format.

Table 3.1: Whisper model size and WER on TED-LIUM v3 and FLEURS (en_us), reproduced from [10].

Model	Params	WER on TED-LIUM3	WER on FLEURS (en_us)
Tiny.en	39M	6.0	11.6
Base.en	74M	4.9	7.6
Small.en	244M	4.0	6.0
Medium.en	769M	4.1	5.0
Large	1550M	4.0	4.4

3.4.2 Speaker-Independent Fine-tuning(SI-FT;B2)

The goal of SI-FT is to adapt Whisper from general-domain ASR to impaired speech by learning speaker-invariant features of domain-specific acoustic and phonetic variability, improving generalization to unseen target speakers.

SI-FT performs supervised sequence-to-sequence fine-tuning with utterance-level inputs x and reference transcripts y. To focus training on pronunciation-related deviations, SI-FT is trained on a filtered subset of utterances with pronunciation issues, excluding those with semantic errors (see Chapter 4). To keep the generation target consistent across conditions, the decoder prompt is fixed via `forced_decoder_ids` with `language=English` and `task=transcribe`.

SI-FT uses full-parameter fine-tuning with a cosine learning-rate schedule, and the best checkpoint is selected based on validation WER. In addition to the impaired speech test set, standard English ASR benchmarks (TED-LIUM v3 and FLEURS) are used for evaluation to monitor whether SI-FT introduces degradation on typical speech.

3.4.3 Speaker-Specific Fine-tuning(SS-FT;B3 and B4)

SS-FT further personalizes the model to a target speaker s to capture speaker-specific pronunciation patterns and improve recognition accuracy on that speaker.

SS-FT corresponds to conditions B3 and B4, which share the same per speaker training procedure and differ only in initialization. In B3, SS-FT is initialized from the SI-FT model (B2), while in B4 it is initialized directly from the Whisper baseline (B1). This design allows us to test whether a better population level model (B2) provides a stronger starting point for personalization.

Because per speaker datasets are small, updating all parameters can lead to unstable training and overfitting. To reduce overfitting and improve training stability under limited data, a partial-freezing strategy is applied. This preserves robust low level acoustic representations while allowing higher layers to adapt to speaker specific patterns. At the end of SS-FT, we obtain one personalized checkpoint per speaker for evaluation and analysis.

3.4.4 LLM-based semantic similarity scoring

Beyond literal matching metrics such as WER, this thesis explores whether improved ASR accuracy also improves semantic comprehensibility. To complement WER, LLMs are used as automatic graders to assess semantic similarity between the ASR hypothesis $\hat{y}$ and the reference transcription y on the SS-FT test set. Prompts containing $(y, \hat{y})$ are constructed, and the LLM is asked to output a similarity score from 1–10 according to a predefined rubric, with at most one decimal place. To reduce randomness, deterministic decoding is used for scoring. LLMs ranging from approximately 0.5B to 14B parameters from the Qwen and LLaMA families are evaluated to study scoring stability and the impact of model size.

To better reflect conversational usage and provide additional context, three consecutive patient utterances are concatenated as one evaluation unit before scoring. To mitigate topical and contextual leakage, adjacent utterances around selected test episodes are removed during data preprocessing (see Chapter 4).

3.4.5 Evaluation

Consistent decoding. The Hugging Face `generate()` pipeline does not exactly replicate the official Whisper decoding behavior, such as token suppression, which can lead to different transcriptions for the same model. To ensure fair and consistent comparisons with the official Whisper pipeline, Hugging Face checkpoints are exported after each epoch, converted to the OpenAI Whisper format using the a publicly available conversion script[1] and decoded with the official OpenAI Whisper implementation to produce $\hat{y}$. This conversion and decoding procedure is applied uniformly to all conditions (B1–B4).

Conversational transcription. When multi-turn audio is transcribed with `whisper.transcribe()` for scoring, cross-segment conditioning is disabled by setting `condition_`

[1] Conversion script: `convert_hf_to_openai.py` (zuazo-forks).

`on_previous_text=False`. This reduces error propagation from earlier segments to later segments and improves comparability when segmentation boundaries vary.

Text normalization. Before computing WER, both the reference transcript y and the hypothesis $\hat{y}$ are normalized to reduce non-content differences. The normalization rules are described in Chapter 5 and are applied consistently across all WER computations.

Randomness control. In some cases, ASR decoding produces repetitive tokens that are not meaningful, which can inflate WER, especially when the evaluation set is small. To reduce this issue, decoding uses a small non-zero temperature (temperature = 0.2). Since stochastic decoding introduces variance, each evaluation is repeated a fixed number of times and the average WER is reported. The concrete decoding and randomness settings are specified in Chapter 5.

Chapter 4

Data

4.1 Datasets

This thesis uses **AphasiaBank** as the primary impaired-speech dataset and **TED-LIUM v3** and **FLEURS** as standard-speech reference datasets [8, 5, 2]. The AphasiaBank data are used for training, validation, and testing of the proposed ASR adaptation pipeline, while TED-LIUM v3 and FLEURS are used as out-of-domain checks to verify that adaptation does not significantly harm general ASR performance.

4.1.1 AphasiaBank

AphasiaBank [8] is the primary dataset of impaired speech used in this thesis. It provides audio/video recordings from speakers with aphasia together with time-aligned transcripts, and the data are organized by collection site and participant. AphasiaBank contains data in 13 languages, including Cantonese, English, and French. All recordings are collected following the AphasiaBank protocol, which standardizes instructions across participants and elicits four discourse genres: *personal narratives*, *picture description*, *story retelling*, and *procedural discourse*. The recordings are transcribed in the CHAT format [7]. The transcripts include linguistic annotations such as morphology, grammar, and discourse codes, providing rich linguistic information. Each utterance is linked to the corresponding segment of the video.

In this work, we focus on the English protocol data. The subset used in our experiments covers 32 institutions and 455 speakers with aphasia, totaling approximately 448 hours of recorded sessions.

To support the proposed two-stage ASR adaptation pipeline, we split the data by speaker to ensure that evaluation speakers are strictly unseen during speaker-independent training. Specifically, we use 90% of the speakers to fine-tune Whisper into a speaker-independent impaired-speech ASR model, and reserve the remaining 10% of the speakers for personal adaptation experiments.

4.1.2 TED-LIUM v3

TED-LIUM v3 [5] is a widely used dataset of English TED talks with aligned transcripts. In this thesis, it is included as a typical-speech reference to provide a comparison point for Whisper-based systems outside the impaired-speech domain. Only the TED-LIUM v3 **test** split is used as an out-of-domain evaluation set to illustrate the gap between recognition on typical speech and recognition on AphasiaBank under the same decoding and evaluation settings.

4.1.3 FLEURS

FLEURS [2] is a multilingual speech dataset designed to support speech research across many languages and accents. In this thesis, the English subset is used as an additional standard-speech reference dataset, complementing TED-LIUM v3 with a different style of typical speech. Compared with TED-LIUM v3, which features long-form public speaking, FLEURS contains shorter and more diverse utterances. This provides a second out-of-domain check that improvements on impaired speech do not come at the cost of generalization to typical English speech.

Similar to TED-LIUM v3, FLEURS is only used for evaluation. Only the English test split is used in the experiments, and WER is reported before and after impaired-speech adaptation.

4.2 Preprocessing

This section describes the preprocessing pipeline used to convert raw conversational recordings into standardized training and evaluation examples for Whisper-based ASR speaker-independent fine-tuning and personal adaptation. The preprocessing has two goals. First, the original recordings are segmented into utterance-level clips and converted into a format suitable for Whisper. Second, transcripts are normalized to reduce label noise caused by annotation symbols and transcription conventions, which is especially important for impaired speech.

4.2.1 Audio preprocessing

AphasiaBank recordings are organized by institution and speaker, and each conversation file contains multi-turn interviewer and patient dialogue. Rather than using full-session audio, this thesis operates on utterance-level segments. Each segment is stored as a separate audio clip and paired with a human reference transcript. Utterances are excluded when timestamps are unavailable or when the filtered transcript is empty.

All audio clips are resampled to 16 kHz and converted to a single-channel waveform to match Whisper's default input setting. Segments with missing or corrupted audio are removed. During training and inference, audio is converted to Whisper log-Mel spectrogram features following the default Whisper preprocessing pipeline.

4.2.2 Transcript preprocessing

Transcripts are extracted from the corpus annotations and converted into plain-text targets for ASR training and evaluation. AphasiaBank transcripts are provided in the CHAT format [7] and contain extensive linguistic and paralinguistic annotations. While useful for clinical analysis, these annotation tokens are not intended ASR targets and, if kept, can introduce non-lexical symbols and inflate WER. Therefore, a deterministic preprocessing pipeline is applied to convert each patient utterance into clean, ASR-oriented reference text.

One common pattern in CHAT is explicit replacement annotation, which links a surface form to an intended word or phrase. When such a pattern is present, the intended text is substituted into the reference transcript. This step helps the reference better represent the speaker's intended lexical content and reduces mismatches due to transcription conventions.

In addition, markers that are not directly related to the speaker's semantic content are removed. These include role tags (e.g., `*PAR:`), timing spans, and other special annotation tokens. Markers indicating retracing or repetition (e.g., `[/]` and `[//]`) and short pause symbols (e.g., `(..)`) are also removed. Filled-pause tokens introduced as annotation (e.g., `&-uh`) and gesture codes (e.g., `&=ges:clothing`) are excluded, since they describe transcript structure rather than stable lexical targets for ASR. Finally, extra whitespace is collapsed to obtain clean sentences.

Table 4.1 shows an example. The role tag `*PAR:` and the timing span at the end of the utterance are removed. The retracing marker `[//]` and the unintelligible token `xxx` are discarded. For pronunciation-related annotations, `@u` denotes a Unibet

Table 4.1: Example of transcript preprocessing from CHAT format to plain text.

Field	Text
Original (CHAT)	`*PAR: and they [//] she was going [//] xxx [//] doʊin@u [: going] [* p:n] there . 440719_447807`
After preprocessing	`and they she was going going there .`

transcription of the produced form, and `[: going]` provides the intended word.The tag `[* p:n]` indicates that the deviation is a pronunciation-related error. Therefore, the Unibet form is replaced with the intended word `going` to better preserve the intended meaning in the reference transcript.

4.3 Data splits

This thesis adopts a two-stage adaptation framework. Therefore, the AphasiaBank data are partitioned in a speaker-disjoint manner into two parts, used for Speaker-Independent Fine-tuning (SI-FT) and Speaker-Specific Fine-tuning (SS-FT), respectively. The SI-FT split is designed to learn cross-speaker pronunciation patterns while reducing semantic label noise, providing a strong cold-start model. At the speaker level, 90% of speakers are assigned to Stage 1 (SI-FT), and the remaining 10% of speakers are reserved for Stage 2 (SS-FT). The SS-FT split is constructed per speaker to evaluate, under strict isolation, how different initialization strategies affect adaptation to an individual speaker. Figure 4.1 provides an overview of these splits.

Speaker-Independent Fine-tuning (SI-FT) Data

Aphasia can involve phonological errors (e.g., phoneme substitutions) as well as semantic errors (e.g., intending "spoon" but producing "fork"). To keep the first stage focused on pronunciation and to reduce label noise from semantic errors, SI-FT uses only utterance-level segments that exhibit pronunciation-related errors while containing no semantic errors. Concretely, an utterance is selected if its CHAT transcript contains at least one pronunciation-related marker of the form `[* p:...]` and contains no semantic-error markers of the form `[* s:...]`.

For the retained utterances, the transcript normalization described in Section 4.2.2 is applied to produce the reference text for ASR training, and the corresponding audio segments are extracted.

4.3. Data splits

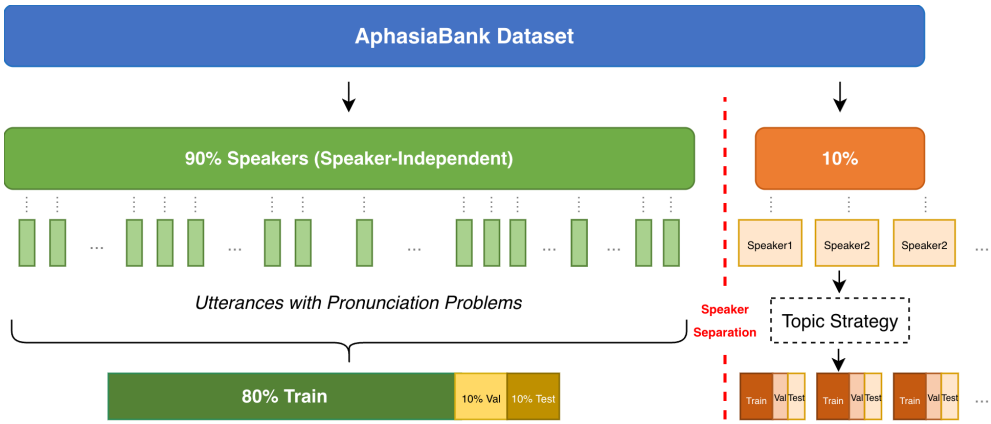


Figure 4.1: Speaker-disjoint data split for the two-stage adaptation framework. AphasiaBank speakers are partitioned into SI-FT (90%) and SS-FT (10%). SI-FT uses utterances with pronunciation errors only and is split into train/validation/test with an 8:1:1 ratio. SS-FT is divided into multiple speaker-specific subsets, and for each speaker, a topic-driven test selection strategy is used to construct speaker-specific train/validation/test splits.

Table 4.2: SI-FT data split statistics after filtering and speaker-disjoint exclusion.

Subset	Utterances	Duration (h)
Train	4284	8.51
Validation	535	1.05
Test	537	1.06
Total	5356	10.62

To prevent speaker leakage between SI-FT and SS-FT, all speakers reserved for SS-FT and their associated recordings are excluded from SI-FT data construction. After this exclusion, the SI-FT dataset contains 5,356 utterances with a total duration of 10.62 hours. The data are then split into training, validation, and test sets using an 8:1:1 ratio. Table 4.2 reports the resulting statistics.

Speaker-Specific Fine-tuning (SS-FT) Data

SS-FT data are constructed separately for each target speaker. Unlike SI-FT, which selects only utterances with pronunciation issues, SS-FT uses all speech from the speaker. Each speaker has an independent train, validation, and test split. Unlike SI-FT, SS-FT evaluation aims to better reflect real communication scenarios. Therefore, the test split is built using a topic-driven selection strategy based on the AphasiaBank protocol, where conversations are organized by topic.

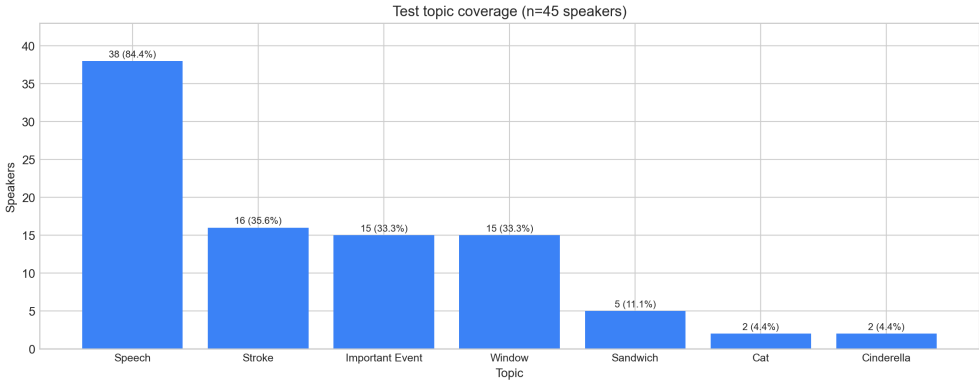


Figure 4.2: Test topic coverage across speakers in the SS-FT split. Speech Topic has the highest prevalence: 84.4% of the 45 speakers have this topic represented in their test sets. Stroke, Important Event, and Window show similar coverage, each at roughly 33.3%. Only a small number of speakers include the Sandwich, Cat, or Cinderella topics.

Because topic coverage differs across speakers, topics are first ordered by overall frequency, more frequent topics are considered earlier. Topics related to picture description are down-weighted in this ranking to emphasize conversational interaction. Next, the test target size is set to 10% of the speaker's available utterances. Starting from the highest-ranked topics, all utterances from a topic are added to the test set. If adding a topic would exceed the target size by a large margin, that topic is skipped. Topic selection stops once the test set reaches or slightly exceeds the target size. This procedure keeps the test size controlled while prioritizing common conversational topics. After applying this selection strategy, the topic coverage distribution is shown in Fig. 4.2. Among all topics, speech has the highest coverage: 84.4% of speakers cover this topic. Meanwhile, the strategy also successfully reduces the proportion of image-description tasks.

After the test set is determined, any utterances within 10 seconds of the test utterances are removed to reduce potential context overlap between training and evaluation. The remaining utterances are split into train and validation sets using a 9:1 ratio, leading to an overall proportion close to 8:1:1.The validation split is used for early stopping and checkpoint selection during SS-FT, while the test split is used only for final reporting.

After splitting, the SS-FT data include 45 speakers and 10,444 utterances, with 11.11 hours of speech in total. The amount of available data varies notably across speakers. Total duration per speaker ranges from 0.023 to 0.835 hours, with a mean of

0.247 hours. This corresponds to an average of 232.1 utterances per speaker. Detailed per-speaker split statistics and the selected test topics are provided in Appendix A.

The SS-FT test split is also used as the evaluation set for LLM-based semantic similarity scoring in RQ3.

4.4 Leakage Prevention

Since this thesis adopts a two-stage adaptation framework and compares different initialization strategies across stages, it is essential to prevent overlap between training and test data. Such leakage can happen at the speaker level, at the utterance level, or through adjacent conversational context, and it can lead to overly optimistic evaluation results. This thesis applies the following strategies.

- **Speaker-level isolation.** Data are split by speaker. All target speakers used for Stage 2 SS-FT are completely unseen during Stage 1 SI-FT. This ensures that no target speaker is seen during SI-FT.
- **No overlap across subsets.** Within each stage, the training, validation, and test sets are mutually exclusive. Each utterance is assigned to exactly one subset and does not appear in any other subset.
- **Context leakage control.** After the SS-FT test set is constructed using the topic-driven selection strategy, all utterances within a 10-second window before or after each test utterance are removed from the remaining pool. This reduces the chance that the training data contain adjacent turns that reveal contextual cues for the test samples.

Chapter 5

Experiments

This chapter describes the experimental setup used throughout the thesis, including the compute environment, training setup, decoding configuration, and checkpoint selection. Unless otherwise stated, the same preprocessing pipeline, dataset splits, decoding configuration, and evaluation protocol are applied across all conditions to ensure fair comparisons.

5.1 Experimental Setup

5.1.1 Compute environment.

All experiments were conducted on NVIDIA A100-SXM4 GPUs with 40 GB memory. Model training was implemented with PyTorch and Hugging Face Transformers, and evaluation decoding was performed with the official OpenAI Whisper implementation to match the reference decoding behavior. Mixed precision training with fp16 was enabled to increase throughput and reduce GPU memory usage.

5.1.2 Training Setup

Both stages adopt supervised sequence to sequence fine tuning of Whisper. Stage 1 performs Speaker-Independent Fine-tuning (SI-FT) on multi-speaker impaired-speech data to obtain a population-level model. Stage 2 performs Speaker-Specific Fine-tuning (SS-FT) on speaker-specific data. In the experimental conditions, B3 initializes SS-FT from the SI-FT model (B2), while B4 initializes SS-FT directly from the Whisper baseline (B1).

Table 5.1: Training hyperparameters used for SI-FT and SS-FT.

Hyperparameter	SI-FT (B2)	SS-FT (B3/B4)
Backbone init.	`whisper-small.en`	from B2 (B3) or from B1 (B4)
Learning rate	7×10^{-6}	3×10^{-6}
Epochs	12	5
Batch size (per device)	8	4
Weight decay	0.01	0.0
Eval batch size	8	4
Frozen encoder layers	none	layers 0–5 frozen

Since SS-FT typically uses much less data than SI-FT, updating all parameters may lead to unstable training and overfitting. A more conservative update strategy is therefore used for SS-FT, including a smaller learning rate, fewer training epochs, and partial encoder freezing. Specifically, for `whisper-small.en`, the first 6 encoder layers out of 12 are frozen, and only the remaining encoder layers and the decoder are updated.

For both stages, the language and task prompts are fixed by setting `forced_decoder_ids` to English and `transcribe`, ensuring consistent generation targets across conditions. Key hyperparameters are summarized in Table 5.1, and the complete hyperparameter listing is provided in Appendix B.

5.1.3 Decoding configuration

To ensure consistent and comparable evaluation across conditions, ASR evaluation uses the same OpenAI Whisper decoding backend as described in Section 3. Hugging Face checkpoints are converted to the OpenAI Whisper format before decoding. Apart from model weights, conditions B1–B4 share the same decoding configuration, with language fixed to English and task fixed to transcription.

For SS-FT evaluation, a small non-zero decoding temperature (temperature = 0.2) is used to reduce occasional degenerate repetition on small test sets. Since temperature > 0 introduces randomness, results under sampling are reported as the average of three independent runs.

Checkpoint Selection

Model selection is performed based on validation set WER computed with the same decoding and text normalization protocol used for final test evaluation. The checkpoint with the lowest validation WER is selected as the final model. This procedure is applied once for SI-FT and independently for each target speaker in SS-FT.

5.2 Baselines

This section summarizes the four experimental conditions compared in this thesis. The conceptual definitions of B1–B4 are introduced in Chapter 3. Here, the conditions are restated in a compact form for experimental reference.

- **B1: Whisper (baseline).** The original `whisper-small.en` model without fine-tuning, provides a zero shot baseline on the target domain.
- **B2: SI-FT.** (fine-tune)Speaker-Independent Fine-tuning on multi-speaker impaired-speech data, producing a population-level adapted model.
- **B3: SI-FT→SS-FT (two stage).** Speaker-Specific Fine-tuning for each target speaker, initialized from the SI-FT model (B2).
- **B4: Whisper→SS-FT (direct adaptation).** Speaker-Specific Fine-tuning for each target speaker, initialized directly from the Whisper baseline (B1).

In terms of research questions, RQ1 evaluates the benefit of SI-FT by comparing B1 and B2. RQ2 evaluates whether SI-FT provides a better initialization for personalization by comparing B3 and B4. RQ3 performs an exploratory LLM-based similarity scoring analysis using ASR outputs produced under B1–B4.

5.3 Experimental protocol

All experiments follow the speaker-disjoint split strategy and leakage prevention rules described in Chapter 4. Across conditions B1–B4, the same audio preprocessing, transcript preprocessing, and text normalization are applied. For ASR evaluation, all models are converted to the OpenAI Whisper format and decoded with the official OpenAI Whisper implementation under the same task configuration. Results are reported on the corresponding test splits using a consistent aggregation method across conditions.

5.4 Metrics

5.4.1 Word Error Rate (WER)

ASR performance is primarily evaluated using word error rate (WER) [6, 9]. Given a reference transcript y and a hypothesis transcript $\hat{y}$, WER measures the minimum number of word-level edit operations required to transform $\hat{y}$ into y, normalized by the reference length. Let S, D, and I denote the numbers of substitutions, deletions, and insertions, and let N denote the number of words in the reference transcript.

WER is defined as

$$\mathrm{WER}(y, \hat{y}) = \frac{S + D + I}{N}. \tag{5.1}$$

WER is reported as a percentage. Since insertions are not bounded by N, WER can exceed 100% when the hypothesis is much longer than the reference. This thesis computes WER using the `jiwer` toolkit.[1]

To avoid counting superficial formatting differences as errors, both y and $\hat{y}$ are normalized with a fixed text-normalization policy before WER computation. The normalization is based on Whisper's English normalizer with additional project-specific rules to handle dataset-specific transcript artifacts. The exact normalization rules are listed in Appendix C.

5.4.2 LLM-based Similarity

WER measures literal word-level agreement, but the downstream goal in this thesis concerns whether improved transcriptions help a language model infer the intended meaning of speech from people with aphasia. To complement WER, an exploratory LLM-based semantic similarity score is used to estimate how well an ASR hypothesis preserves the meaning of the reference transcript.

For each evaluation unit, the scoring model is given a reference record r and an ASR hypothesis $\hat{y}$, and outputs a scalar score $s(y^{CHAT}, \hat{y}) \in [0, 10]$, where higher values indicate closer meaning. Here, y^{CHAT} is taken from the original CHAT-style record, prior to text cleanup, and may include additional contextual cues such as non-verbal annotations (e.g., gestures). The rubric instructs the model to focus on meaning and to ignore grammar and minor word errors. The model is required to output a single numeric score with at most one decimal place, which is parsed automatically. The prompt template are provided in Appendix C.1.

[1] `https://jitsi.github.io/jiwer/`

Scoring is performed at chunk level to provide short-span discourse context rather than scoring isolated utterances. Every three consecutive utterances are concatenated into one chunk, and the reference text is chunked using the same rule. This yields paired sequences $(r_1, \ldots, r_K)$ and $(h_1, \ldots, h_K)$ for reference chunks and ASR chunks. Each pair (r_k, h_k) is scored to obtain a sequence of scores $\{s_k\}_{k=1}^{K}$.

Comparability across ASR conditions is enforced via a single-turn scoring protocol. For each chunk index k, the scoring model receives only the current chunk pair (r_k, h_k) and returns a similarity score.

To assess stability across model families and sizes, scoring is repeated with multiple LLMs from the Qwen and LLaMA families.

- Qwen: `Qwen2.5-0.5B-Instruct`, `Qwen2.5-3B-Instruct`, `Qwen2.5-7B-Instruct`, `Qwen2.5-14B-Instruct`.

- LLaMA: `Llama-3.2-1B-Instruct`, `Llama-3.2-3B-Instruct`, `Llama-3.1-8B-Instruct`.

Deterministic decoding is used for scoring (temperature = 0), and the same prompt and output format are used across all models and conditions.

Scores are aggregated at the speaker level using a two-level procedure aligned with the topic-based test construction. For each speaker, chunk scores are first averaged within each topic, and the speaker-level score is then computed by averaging these topic-level means across all test topics for that speaker. The overall score is reported as the mean across speakers (macro average), which avoids over-weighting speakers with more test data.

$$\bar{s} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(\frac{1}{|\mathcal{T}_s|} \sum_{t \in \mathcal{T}_s} \left(\frac{1}{K_{s,t}} \sum_{k=1}^{K_{s,t}} s_{s,t,k} \right) \right), \tag{5.2}$$

where $\mathcal{S}$ is the set of test speakers, $\mathcal{T}_s$ is the set of test topics for speaker s, $K_{s,t}$ is the number of chunks in topic t, and $s_{s,t,k}$ is the LLM score for chunk k.

5.5 RQs experiments

This section defines the concrete comparisons used to answer RQ1–RQ3 under the shared protocol and metrics described above.

5.5.1 RQ1: SI-FT vs Whisper baseline

RQ1 tests whether speaker-independent fine-tuning improves ASR on impaired speech. The comparison is performed between B1 (Whisper baseline) and B2 (SI-FT). Both conditions are evaluated on the SI-FT test split described in Section 4.3. In addition, TED-LIUM v3 and FLEURS (en_us) test splits are used as out-of-domain checks under the same OpenAI Whisper decoding backend and text-normalization policy. Results are reported as overall WER.

5.5.2 RQ2: SI-FT→SS-FT vs Whisper→SS-FT

RQ2 evaluates whether SI-FT provides a better initialization for speaker-specific fine-tuning. The primary comparison is between B3 (SI-FT→SS-FT) and B4 (Whisper→SS-FT), which use the same SS-FT training data and training procedure and differ only in initialization. Evaluation is performed on the SS-FT test split for each target speaker, and results are aggregated at the speaker level. For interpretability, B1 and B2 are also evaluated on the same SS-FT test split as non-personalized reference points.

For SS-FT evaluation, decoding for WER uses temperature = 0.2 to mitigate occasional degenerate repetitions in the hypotheses. Since temperature > 0 introduces randomness, decoding is repeated three times and results are reported as the average.

5.5.3 RQ3: LLM-based similarity scoring

RQ3 explores whether improved ASR transcription quality leads to improved semantic fidelity. LLM-based similarity scoring is computed on the SS-FT test split using the protocol defined in Section 5.4.2. The comparison factors are:

- **ASR condition**: B1–B4.
- **LLM scale**: Qwen (0.5B, 3B, 7B, 14B) and LLaMA (1B, 3B, 8B).

Scores are aggregated by topic and speaker, and trends are compared across ASR conditions and LLM sizes.

Chapter 6

Results and Discussion

6.1 RQ1: SI-FT vs Whisper baseline

RQ1 investigates whether Speaker-Independent Fine-tuning (SI-FT) can significantly reduce recognition error rates on impaired speech compared with the Whisper baseline. It also discusses the trade-off between in-domain gains and potential degradation on out-of-domain benchmarks. Following the experimental setup and methodology described in the previous chapters, the overall results are summarized in Table 6.1.

Table 6.1: WER (%) for RQ1 (B1 vs. B2). After fine-tuning, WER on the SI-FT test set drops from 47.38 to 29.30. On the standard-speech datasets, WER changes only marginally: FLEURS increases slightly (+1.21), while TED-LIUM v3 decreases by 0.33.

Dataset	**B1**	**B2**
SI-FT test (AphasiaBank)	47.38	29.30
TED-LIUM v3	4.01	3.68
FLEURS en_us	6.00	7.21

Before adaptation, the vanilla Whisper baseline achieves a WER of 47.38% on the SI-FT test set, which is substantially higher than its WER on typical-speech benchmarks, 4.01% on TED-LIUM v3 and 6.00% on FLEURS. This highlights that impaired speech is significantly harder for general-purpose ASR models and that off-the-shelf recognition can be unreliable for this population. At the same time, the typical-speech results are consistent with the performance reported for Whisper on standard English data, which supports the correctness of the evaluation pipeline and

experimental configuration.

After speaker-independent fine-tuning, WER on the SI-FT test set decreases from 47.38% to 29.30%. This corresponds to an absolute reduction of 18.08 WER points and a relative reduction of approximately 38.2%, which is illustrated in Figure 6.1. These results indicate that speaker-independent adaptation using multi-speaker impaired-speech data can substantially improve in-domain recognition performance, confirming the effectiveness of SI-FT for RQ1.

The out-of-domain checks show mixed but overall controlled effects. On TED-LIUM v3, WER decreases slightly from 4.01% to 3.68%, indicating a small improvement rather than degradation. On FLEURS, WER increases from 6.00% to 7.21%, reflecting a mild degradation. This pattern is consistent with the expectation that fine-tuning toward impaired-speech acoustic and linguistic distributions can shift model parameters toward the target domain, which may reduce performance on some typical-speech benchmarks. However, the magnitude of degradation on FLEURS is small in absolute terms, while the gain on the impaired-speech test set is large. Moreover, WER remains low on both typical-speech test sets, suggesting that out-of-domain degradation is limited in this setting. Overall, the results indicate that careful selection of fine-tuning data can avoid severe degradation and may even improve typical-speech performance on some benchmarks without explicitly mixing typical speech during training. Therefore, SI-FT is supported as a stronger cold-start model for impaired speech, and the remaining out-of-domain effects appear controllable.

Beyond validating the effectiveness of speaker-independent adaptation, the RQ1 findings also motivate RQ2. If SI-FT learns population-level acoustic and pronunciation patterns closer to impaired speech, it may serve as a better initialization for subsequent speaker-specific fine-tuning, potentially improving the ceiling or stability of personalization.

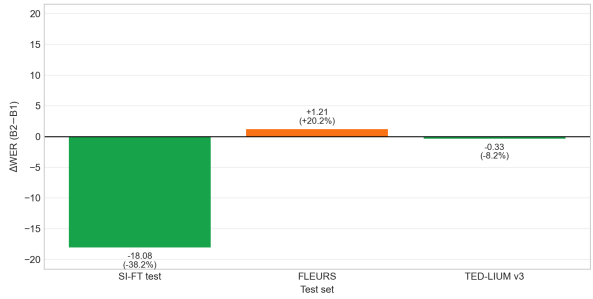


Figure 6.1: WER change for RQ1 (ΔWER = WER(B2: SI-FT) − WER(B1: Whisper)) across datasets. SI-FT reduces WER on the SI-FT test set by 38.2% (relative), while changes on FLEURS and TED-LIUM v3 are small.

6.2 RQ2: SI-FT→SS-FT vs Whisper→SS-FT

RQ2 examines whether speaker-independent fine-tuning (SI-FT) provides a better initialization for subsequent speaker-specific fine-tuning. The core comparison in this section is between B3 (SI-FT→SS-FT) and B4 (Whisper→SS-FT), with B1 and B2 reported as reference points. Following the setup in Section 5, SS-FT-related conditions are decoded with temperature = 0.2 and decoded three times to reduce variance introduced by stochastic decoding. The overall results for B3 and B4 are therefore summarized using the mean over three decoding runs.

Table 6.2 summarizes overall WER on the SS-FT test split across the three decoding runs. B2 achieves an average WER of 32.64 across the 45 speakers, which is substantially lower than B1 at 46.13. This further supports the RQ1 finding that SI-FT provides strong cold-start benefits. In contrast, the mean WER of B3 is slightly higher than B2, while the mean WER of B4 is substantially higher than B1. B3 remains close to B2 at the overall level, while B4 is markedly worse and also exhibits larger run-to-run values.

To better understand these results, Figure 6.2 shows the per-speaker WER distributions. The boxplot indicates that extreme WER outliers appear in B1, B3, and B4. Manual inspection indicates that these outliers are often associated with degenerate repetitive outputs during decoding, which occur more often in utterances with longer silent intervals. The boxplot also indicates that, although the mean of B3 can be slightly higher than B2 due to outliers, the median WER of B3 is lower than B2 from 23.3% to 22.5%, suggesting that personalization on top of SI-FT can still provide consistent small gains for many speakers. In contrast, B4 is consistently worse than

Table 6.2: Overall WER (%) on the SS-FT test set, reported as macro average over speakers. For each baseline, decoding is repeated three times (temperature = 0.2) and the table reports mean ± standard deviation across runs.

Condition	Overall WER (%, mean ± std)
B1 (Whisper)	46.13 ± 0.00
B2 (SI-FT)	32.64 ± 0.00
B3 (SI-FT→SS-FT)	33.03 ± 0.93
B4 (Whisper→SS-FT)	50.38 ± 1.94

B3, both in median (25.4%) and mean (50.4%).

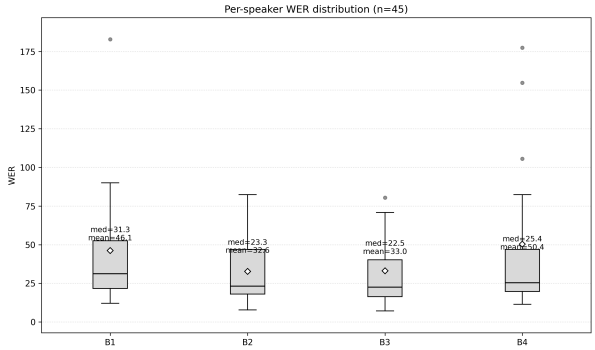


Figure 6.2: Per-speaker WER distributions for RQ2 under four conditions (B1–B4; $n = 45$). Mean WERs are higher in B1 (46.1) and B4 (50.4) than in B2 (32.6) and B3 (33.0). The median WER is highest for B1 (31.3), followed by B4 (25.4); B2 (23.3) is slightly higher than B3 (22.5). Outliers are observed in B1, B3, and B4, with B4 showing the most pronounced outliers.

To directly test the impact of initialization, Figure 6.3 shows a paired per-speaker comparison between B3 and B4. For the majority of speakers, B3 yields lower WER than B4. Among the 45 target speakers, 37 satisfy $\mathrm{WER}_{\mathrm{B3}} < \mathrm{WER}_{\mathrm{B4}}$, corresponding to 82.22%. This result suggests that the population-level acoustic and pronunciation deviations learned by SI-FT serve as a useful prior for SS-FT, making personalization more reliable for most speakers. A small number of speakers still perform better under B4, which suggests that personalization gains can depend on speaker-specific factors such as speech patterns, available data quantity, and topic coverage, and that small-sample fine-tuning can remain unstable in some cases.

are largely within ±10 WER points. Figure 6.4 further compares B2 and B3 at the speaker level. In 45 speakers, the proportion for which B3 further reduces WER com-

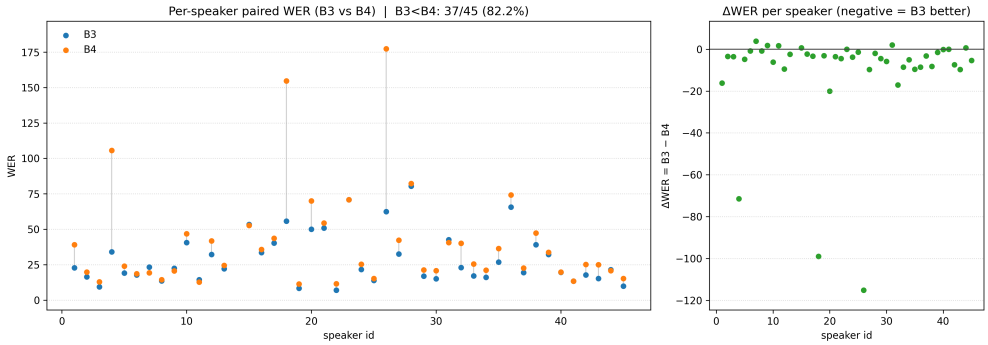


Figure 6.3: Paired per-speaker WER comparison between conditions B3 and B4 ($n = 45$). Left: pairedd per-speaker WERs for B3 and B4. Right: per-speaker difference ΔWER = B3 − B4 (negative values indicate B3 performs better). B3 achieves lower WER than B4 for 37/45 speakers (82.2%). Most differences are below 0 and concentrated near 0, with the majority many are above -10, while a few speakers show substantially larger improvements.

pared to B2 is 65.9%. This indicates that although SI-FT already provides substantial cold-start improvements, SS-FT can still bring additional gains for many individuals. However, compared with the clear advantage of B3 over B4, the improvement of B3 over B2 is less pronounced. One possible reason is that the pronunciation deviations of some speakers are already well captured by SI-FT, leaving limited room for SS-FT to improve further. In addition, SS-FT uses all utterances from a speaker, which may introduce label noise related to language-level symptoms beyond pronunciation, limiting the potential of personalization.

Per-speaker results under all four baselines, averaged over the three decoding runs, are provided in Appendix D.

Overall, RQ2 supports the main claim that a two-stage strategy, SI-FT followed by SS-FT, is typically better and more stable than personal adaptation starting directly from the Whisper baseline. The results also suggest that, compared to B1 and B4, degenerate repetitions occur less frequently under B2 and B3. These findings provide the basis for RQ3, which examines whether lower WER under improved ASR conditions is accompanied by higher downstream semantic similarity scores.

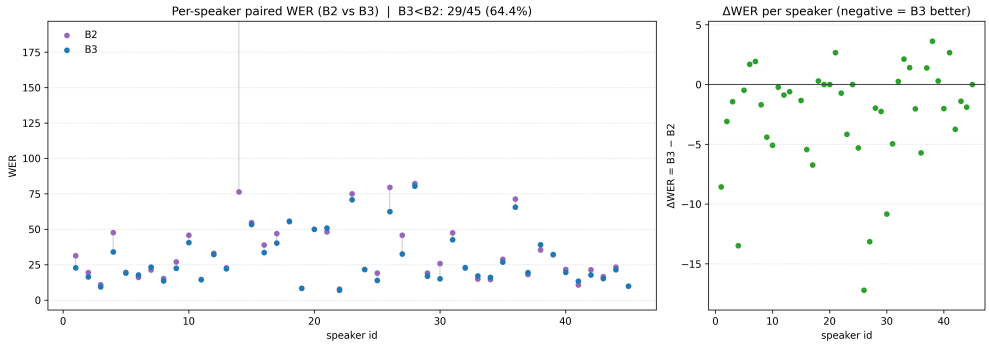


Figure 6.4: Paired per-speaker WER comparison between conditions B2 and B3 ($n = 45$). Left: paired per-speaker WERs for B2 and B3. Right: per-speaker difference $\Delta\text{WER} = \text{B3} - \text{B2}$ (negative values indicate B3 performs better). Although B3 yields lower WER than B2 for 29/45 speakers (64.4%), the magnitude of change is generally small. Most ΔWER values lie above 0, and the differences between the two baselines are largely within ± 10 WER points.

6.3 RQ3: LLM-based similarity scoring

RQ3 explores whether improvements in ASR quality translate into better semantic-level comprehensibility, enabling downstream models to more accurately recover the speaker's intended meaning. Since no human semantic annotations are available, this thesis uses the LLM-based semantic similarity score defined in Section 5.4.2 as a proxy metric. On the SS-FT test set, the reference text y and the ASR hypothesis $\hat{y}$ are aligned and scored at the chunk level. Scores are then aggregated by averaging within each topic, averaging across topics for each speaker, and finally taking a macro average across speakers.

Table 6.3: RQ3 overall LLM-based similarity scores for the Qwen family.

Model	B1	B2	B3	B4
0.5B	0.002	0.003	0.002	0.006
3B	0.781	0.514	0.632	0.805
7B	4.529	4.583	4.701	4.764
14B	5.190	5.426	5.523	5.405

At the overall level, the mean similarity scores across ASR conditions (B1–B4) differ only slightly, as shown in Tables 6.3 and 6.4. The results also show that similarity scores generally increase with model size. However, for a fixed scoring model, the score

differences across training strategies are not pronounced. This suggests that under the current scoring protocol and scale, overall similarity scores have limited sensitivity to distinguish different ASR training strategies. Therefore, the analysis for RQ3 focuses on per-speaker relationships, which better capture individual variability.

Table 6.4: RQ3 overall LLM-based similarity scores for the LLaMA family.

Model	B1	B2	B3	B4
1B	1.629	1.621	1.604	1.581
3B	5.995	6.016	6.057	6.111
8B	3.626	3.437	3.604	3.689

Despite small differences in overall means, a negative association is observed between per-speaker WER and LLM-based similarity scores. Table 6.5 summarizes Spearman correlations across the four ASR conditions. For Qwen2.5-14B, correlations consistently fall between -0.86 and -0.87. For LLaMA-3.1-8B, correlations are also consistently negative, ranging approximately from -0.77 to -0.85. These results indicate that, across speakers, lower WER typically corresponds to higher semantic similarity scores, meaning that more accurate transcripts are more helpful for downstream models to recover intended meaning.

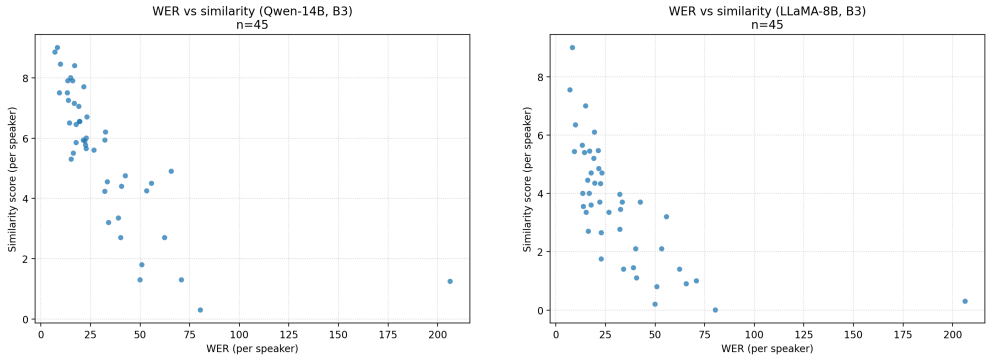
Table 6.5: Spearman correlation between per-speaker WER and LLM-based similarity score. Negative values indicate that lower WER is associated with higher similarity.

ASR condition	Qwen2.5-14B	LLaMA-3.1-8B
B1	-0.8737	-0.8296
B2	-0.8710	-0.7748
B3	-0.8689	-0.8103
B4	-0.8641	-0.8519

To visualize this relationship, Qwen2.5-14B and LLaMA-3.1-8B are used as illustrative examples. Figures 6.5a and 6.5b plot per-speaker WER against per-speaker similarity scores, using B3 as an example ASR condition. Both judges show a clear monotonic trend and negative correlation, further supporting the intuition that improved ASR quality benefits downstream semantic understanding.

The plots suggest that a small number of extreme WER values substantially inflate the mean differences between conditions. In contrast, the LLM rating uses a bounded 1–10 scale, which can compress large performance gaps and may partly explain why

6.3. RQ3: LLM-based similarity scoring



(a) Qwen **(b)** Llama 8B

Figure 6.5: Per-speaker WER plotted against per-speaker similarity score under condition B3. Left: Qwen-14B. Right: LLaMA-8B. Both plots show a negative association: as WER decreases, similarity scores increase. For Qwen-14B, speakers with WER below 25 typically obtain scores of 6 or higher, whereas the LLaMA-8B scores are more dispersed.

the baselines appear less separable at the scoring stage.

Table 6.6: Utterance-level mean WER after outlier speaker removal.

Baseline	Mean WER
B1 (Whisper)	39.978960
B2 (SI-FT)	35.845197
B3 (SI-FT→SS-FT)	32.515670
B4 (Whisper→SS-FT)	34.053274

To reduce the influence of extreme cases, speakers were filtered by removing those for whom at least one baseline's WER exceeded another baseline by more than 100 WER points. This procedure removed three speakers. After filtering, utterance-level mean WERs were recomputed for all baselines; the results are reported in Table 6.6.

After outlier removal, the worst-performing baseline remains Baseline 1 (39.98), while the best-performing system is Baseline 3 (32.51). However, the differences across the four baselines remain relatively modest. One plausible explanation is limited separation at the overall level due to insufficient test data for some speakers, which makes it difficult to reveal differences between baselines. In this dataset, 19 out of 45 speakers have fewer than 20 test samples. In addition, although WER differs across systems, the transcripts may still capture the key content correctly, resulting in a smaller impact on semantic similarity. Moreover, because this is an automatic proxy metric, it may be affected by judge bias, scale design, and prompt constraints.

In summary, the exploratory RQ3 experiments show a negative correlation between WER and LLM-based semantic similarity at the speaker level. This supports the end-to-end motivation of this thesis and motivates future work to further investigate linguistic information in the ASR outputs, such as part-of-speech patterns, and to incorporate weighted scoring schemes that assign higher importance to semantically critical words, thereby better reflecting how recognition quality affects semantic understanding.

Chapter 7

Conclusions

7.1 Findings

This thesis addresses ASR adaptation for aphasia-related disordered speech. A two-stage adaptation framework based on Whisper is proposed and evaluated, consisting of Speaker-Independent Fine-tuning (SI-FT) and Speaker-Specific Fine-tuning (SS-FT). Four experimental conditions (B1–B4) are constructed to analyze adaptation gains and the effect of initialization. In addition, an exploratory LLM-based semantic similarity scoring protocol is used to examine whether ASR improvements are associated with improved semantic interpretability. Experiments are organized around three research questions, leading to the following conclusions.

RQ1: Can speaker-independent adaptation significantly improve ASR on impaired speech? The results show that SI-FT substantially reduces WER on the processed AphasiaBank test set compared with the unadapted Whisper baseline. This indicates that speaker-independent adaptation using multi-speaker impaired-speech data can effectively reduce distribution mismatch between general-purpose ASR and disordered speech, providing a stronger cold-start model. Out-of-domain evaluation on TED-LIUM v3 and FLEURS suggests that changes in typical-speech performance remain largely controllable under appropriate data selection and fine-tuning settings, and that adaptation does not necessarily cause severe degradation.

RQ2: Does speaker-independent adaptation provide a better initialization for personalization? In the personalization stage, the two-stage path SI-FT→SS-

FT outperforms direct speaker-specific fine-tuning from the Whisper baseline for most speakers. Speaker-level comparisons show that B3 achieves lower WER than the direct path for the large majority of target speakers, which supports the interpretation that SI-FT learns useful population-level priors that provide a better starting point for SS-FT and make personalization more stable and consistent. At the same time, a small subset of speakers performs better under direct personalization, indicating that individual differences can still have a strong influence on personalization outcomes.

RQ3: Do ASR improvements translate into improved semantic-level interpretability? Without human semantic annotations, this thesis uses LLM-based semantic similarity scoring as an exploratory proxy metric. At the overall mean level, score differences across ASR conditions are small, suggesting limited sensitivity of the current scoring protocol and scale for distinguishing training strategies. Therefore, the current results do not allow concluding that ASR improvements can narrow the performance gap between smaller and larger LLMs. However, at the per-speaker level, WER and semantic similarity scores exhibit a negative correlation, indicating that more accurate transcripts are typically associated with higher semantic consistency. This finding supports the end-to-end motivation of the thesis, namely that improving ASR quality has the potential to improve downstream recovery of the speaker's intended meaning.

7.2 Limitations

Despite efforts to control experimental conditions, this thesis has the following limitations.

Limited and imbalanced SS-FT data. In Stage 2, speaker-specific data are limited overall, and the available duration varies substantially across speakers. This constrains the ceiling of personalization and can lead to unstable training or inconsistent gains for some individuals, which may affect the generalizability of conclusions.

Label noise in conversational transcripts. Reference transcripts are obtained from human transcription of conversational speech. Compared to scripted read-speech data with highly controlled labels, conversational speech is harder to transcribe accurately and may contain errors or inconsistencies. Such label noise affects the upper

bound and interpretability of WER evaluation and may also influence downstream semantic scoring.

RQ3 relies on an exploratory proxy metric. The semantic similarity score relies on an LLM acting as an automatic rater. Although stable speaker-level correlations are observed, the score can still be affected by rater bias, scale resolution, and prompt design. It provides trend-level evidence but does not replace human evaluation.

7.3 Future work

Based on the above limitations and experimental observations, future work can proceed in the following directions.

Investigating the impact of part-of-speech information on LLM semantic understanding. To further examine how ASR transcripts affect LLMs' semantic understanding, part-of-speech information in the recognized outputs can be analyzed to identify which types of recognition improvements contribute most to semantic gains and which types of errors are most detrimental. Assigning higher weights to word categories that carry more semantic content may better capture the influence of ASR quality on downstream LLM performance and improve the effectiveness of ASR outputs for subsequent generation.

Scaling and improving personalization data collection. More systematic data collection protocols could increase the amount of speaker-specific training data and reduce variability across individuals. Higher-quality reference text could also be obtained by combining automatic transcription with user confirmation in short interactive dialogues, which may be more practical than fully manual transcription while remaining closer to conversational use.

More rigorous downstream evaluation and interactive tasks. Future work can incorporate human semantic evaluation or task-based metrics. This would enable a stronger validation of whether ASR improvements reliably translate into better downstream understanding and interaction quality. In addition, further study is needed on robustness and cost trade-offs across LLM sizes, and on the deployment boundary where smaller models become viable under a stronger ASR front-end.

In summary, this thesis validates the effectiveness of a two-stage Whisper adaptation framework for impaired speech ASR and demonstrates the advantage of SI-FT as an initialization for personalization. It also provides exploratory evidence linking ASR accuracy to downstream semantic similarity. With larger datasets, finer-grained subgroup analysis, and more reliable evaluation protocols, future work can further support real-world deployment in assistive communication settings.

Appendix A

SS-FT Dataset Details

This appendix provides additional details on the per-speaker split statistics for the SS-FT dataset.

Table A.1: Personal adaptation (SS-FT) data statistics per speaker.

Institution	Person	Train	Val	Test	Hours	Selected topics
ACWT	ACWT01	54	5	9	0.08	Speech, Stroke
ACWT	ACWT12	203	22	28	0.27	Important Event, Sandwich, Speech
APROCSA	aprocsa1713a	139	15	18	0.18	Sandwich, Speech, Window
APROCSA	aprocsa1731a	284	31	52	0.15	Cinderella, Speech, Window
APROCSA	aprocsa1944a	198	22	29	0.31	Important Event, Speech
Adler	adler05	169	18	31	0.19	Speech, Stroke
Adler	adler22	115	12	15	0.14	Important Event, Speech
BU	BU03	61	6	12	0.16	Important Event, Speech
Baycrest	Baycrest12288a	158	17	23	0.19	Cat, Speech, Window
CMU	cmu02	233	25	30	0.14	Cinderella, Important Event, Speech
Elman	elman07	178	19	33	0.16	Important Event, Speech
Fridriksson	fridriksson01	226	25	30	0.24	Sandwich, Speech, Window
Fridriksson	fridriksson05	120	13	19	0.43	Important Event, Speech
Fridriksson	fridriksson06	276	30	35	0.25	Speech, Stroke
Fridriksson	fridriksson09	250	27	34	0.28	Speech, Stroke
Fridriksson	fridriksson12	200	22	42	0.16	Important Event, Speech
Fridriksson-2	1035	738	81	89	0.70	Sandwich, Window
Fridriksson-2	1082	363	40	48	0.77	Window
Fridriksson-2	1097	25	2	6	0.09	Window

(continued on next page)

Institution	Person	Train	Val	Test	Hours	Selected topics
Fridriksson-2	1112	73	8	9	0.25	Window
Garrett	garrett01	198	22	20	0.11	Stroke
Kansas	kansas04	96	10	13	0.16	Important Event, Speech
Kansas	kansas09	41	4	10	0.02	Speech, Stroke
NEURAL	68	149	16	17	0.27	Speech, Stroke
NEURAL-2	110	511	56	69	0.47	Speech, Stroke
NEURAL-2	197	211	23	28	0.16	Speech, Stroke
NEURAL-2	199	315	34	49	0.73	Speech, Stroke
Richardson	richardson14	44	4	5	0.05	Window
SCALE	scale20	110	12	11	0.14	Speech, Stroke
TCU-bi	tcu10	216	24	22	0.27	Important Event, Speech
Thompson	thompson09	88	9	19	0.15	Important Event, Speech
Thompson	thompson11	369	41	86	0.83	Speech, Stroke
Tucson	tucson10	72	8	12	0.19	Speech, Stroke
Tucson	tucson20	213	23	27	0.27	Important Event, Window
Tucson	tucson22	90	10	20	0.08	Speech, Stroke
Whiteside	whiteside12	111	12	20	0.06	Speech, Window
Whiteside	whiteside13	72	7	17	0.14	Speech, Stroke
Williamson	williamson01	103	11	20	0.12	Speech, Window
Williamson	williamson07	267	29	43	0.50	Cat, Speech, Window
Williamson	williamson09	342	38	59	0.37	Speech, Stroke
Williamson	williamson10	101	11	11	0.12	Speech, Window
Williamson	williamson17	88	9	17	0.11	Important Event, Speech
Williamson	williamson24	115	12	27	0.21	Important Event, Speech
Wozniak	wozniak01	174	19	23	0.28	Sandwich, Speech, Window
Wright	wright202a	130	14	20	0.15	Important Event, Speech

Appendix B

Hyperparameters

Table B.1: Training hyperparameters used for SI-FT and SS-FT.

Hyperparameter	SI-FT (B2)	SS-FT (B3/B4)
Backbone init.	`whisper-small.en`	from B2 (B3) or from B1 (B4)
Trainer	`Seq2SeqTrainer`	`Seq2SeqTrainer`
Optimizer	`AdamW`	`AdamW`
Learning rate	7×10^{-6}	3×10^{-6}
LR scheduler	cosine	cosine
Warmup	warmup_ratio = 0.1	warmup_ratio = 0.1
Epochs	12	5
Batch size (per device)	8	4
Grad accumulation	1	1
Weight decay	0.01	0.0
Max grad norm	1.0	1.0
Mixed precision	fp16	fp16
Eval strategy	per epoch	per epoch
Eval batch size	8	4
Gen max length	32	55
Best model metric	`eval_wer`	`eval_wer`
Load best at end	True	True
Save strategy	per epoch	per epoch
Random seed	42	42
Frozen encoder layers	none	layers 0–5 frozen

Appendix C

Text normalization for WER

Before computing WER, both the reference transcript y and the hypothesis transcript $\hat{y}$ are normalized using a fixed rule-based pipeline. The pipeline first applies Whisper's official English normalizer (`whisper.normalizers.EnglishTextNormalizer`) and then applies a small set of additional substitutions to reduce WER inflation from common colloquial spellings that are not covered by the base normalizer. Table C.1 lists the exact rules.

Rule (case-insensitive)	Replacement
bhafta	have to
bhadta	had to
bhasta	has to

Table C.1: Additional text normalization rules applied after Whisper's English normalizer.

The same normalization is applied to y and $\hat{y}$ in all experiments.

C.1 LLM-Based Similarity Evaluation Prompt

This appendix documents the prompt template used for the LLM-based similarity evaluation.

C.1.1 Prompt Template

```
You are an SLP working with a person with aphasia.

Task: rate how close the ASR text is in meaning to the patient's original
↪  intended text.

Scale: 0.0 (meaning is very different) to 10.0 (meaning is the same).

Output exactly 1 line:
score: X.Y

Score must have exactly 1 decimal place.
Focus on meaning; ignore grammar and minor word errors.
If ASR text is empty but reference is not, output score: 0.0
No other text.
```

C.1.2 Prompt Construction

The final prompt is constructed by concatenating the scoring system prompt and a context line:

```
SYSTEM_PROMPT_SCORE + "\n\n" + f'Context: The patient is answering: "{intro}"'
```

C.1.3 Definition of `intro`

The variable `intro` corresponds to the protocol-defined topic lead-in from AphasiaBank. The full set of lead-in prompts is available on the AphasiaBank protocol page: TalkBank AphasiaBank instructions (English).

Appendix D

Per-speaker results for RQ2

Table D.1: Per-speaker mean WER over three runs for RQ2 (B1–B4).

institution	person	B1	B2	B3	B4	Δ (B3−B4)
ACWT	ACWT01	48.57	31.43	22.86	39.05	−16.19
ACWT	ACWT12	19.91	19.44	16.36	19.75	−3.40
APROCSA	aprocsa1713a	14.39	10.79	9.35	12.95	−3.60
APROCSA	aprocsa1731a	58.73	47.62	34.13	105.56	−71.43
APROCSA	aprocsa1944a	28.23	19.62	19.14	23.92	−4.78
Adler	adler05	21.61	16.10	17.80	18.64	−0.85
Adler	adler22	20.65	21.29	23.23	19.35	3.87
BU	BU03	16.10	15.25	13.56	14.41	−0.85
Baycrest	Baycrest12288a	31.32	26.92	22.53	20.70	1.83
CMU	cmu02	52.54	45.76	40.68	46.89	−6.21
Elman	elman07	15.92	14.65	14.44	12.74	1.70
Fridriksson	fridriksson01	43.05	33.11	32.23	41.72	−9.49
Fridriksson	fridriksson05	27.54	22.75	22.16	24.55	−2.40
Fridriksson	fridriksson06	298.65	76.35	206.31	516.22	−309.91
Fridriksson	fridriksson09	56.67	54.67	53.33	52.67	0.67
Fridriksson	fridriksson12	41.63	38.91	33.48	35.75	−2.26
Fridriksson-2	1035	67.17	46.97	40.24	43.60	−3.37
Fridriksson-2	1082	63.18	55.45	55.76	154.70	−98.94

(continued on next page)

D.0. LLM-Based Similarity Evaluation Prompt

institution	person	B1	B2	B3	B4	Δ (B3−B4)
Fridriksson-2	1097	12.50	8.33	8.33	11.46	−3.13
Fridriksson-2	1112	90.00	50.00	50.00	70.00	−20.00
Garrett	garrett01	50.00	48.21	50.89	54.46	−3.57
Kansas	kansas04	12.06	7.80	7.09	11.58	−4.49
Kansas	kansas09	70.83	75.00	70.83	70.83	0.00
NEURAL	68	33.58	21.64	21.64	25.37	−3.73
NEURAL-2	110	16.17	19.16	13.87	15.27	−1.40
NEURAL-2	197	182.80	79.57	62.37	177.42	−115.05
NEURAL-2	199	51.57	45.74	32.59	42.30	−9.72
Richardson	richardson14	76.47	82.35	80.39	82.35	−1.96
SCALE	scale20	24.72	19.10	16.85	21.35	−4.49
TCU-bi	tcu10	19.17	25.83	15.00	20.83	−5.83
Thompson	thompson09	52.48	47.52	42.57	40.59	1.98
Thompson	thompson11	45.94	22.62	22.89	40.02	−17.13
Tucson	tucson10	23.40	14.89	17.02	25.53	−8.51
Tucson	tucson20	28.30	14.62	16.04	21.07	−5.03
Tucson	tucson22	31.82	28.79	26.77	36.36	−9.60
Whiteside	whiteside12	74.29	71.43	65.71	74.29	−8.57
Whiteside	whiteside13	22.22	18.06	19.44	22.69	−3.24
Williamson	williamson01	50.91	35.45	39.09	47.27	−8.18
Williamson	williamson07	41.10	31.96	32.27	33.79	−1.52
Williamson	williamson09	24.92	21.62	19.62	19.72	−0.10
Williamson	williamson10	16.00	10.67	13.33	13.33	0.00
Williamson	williamson17	29.91	21.50	17.76	25.23	−7.48
Williamson	williamson24	25.00	16.67	15.28	25.00	−9.72
Wozniak	wozniak01	25.16	23.27	21.38	20.75	0.63
Wright	wright202a	18.52	9.88	9.88	15.23	−5.35

Bibliography

[1] Chen Bao, Chuanbing Huo, Qinyu Chen, and Chang Gao. As-asr: A lightweight framework for aphasia-specific automatic speech recognition, 2025.

[2] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*, 2022.

[3] Google Research. Project relate. `https://sites.research.google/relate/`. Accessed: 2026-01-22.

[4] Jordan Green, Robert MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn Ladewig, Jimmy Tobin, Michael Brenner, Philip Nelson, and Katrin Tomanek. Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. pages 4778–4782, 08 2021.

[5] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. *CoRR*, abs/1805.04699, 2018.

[6] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, NJ, 2001.

[7] Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3 edition, 2000.

[8] Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307, 2011. PMID: 22923879.

[9] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukás Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, et al. The Kaldi speech recognition toolkit. In *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE Signal Processing Society, 2011.

[10] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[11] Gema Rubio-Carbonero. Communication in persons with acquired speech impairment: The role of family as language brokers. *Journal of Linguistic Anthropology*, 32(1):161–181, 2022.

[12] Giulia Sanguedolce, Sophie Brook, Dragos Gruia, Patrick Naylor, and Fatemeh Geranmayeh. When whisper listens to aphasia: Advancing robust post-stroke speech recognition. pages 1995–1999, 09 2024.

[13] Jimmy Tobin, Phillip Nelson, Bob MacDonald, Rus Heywood, Richard Cave, Katie Seaver, Antoine Desjardins, Pan-Pan Jiang, and Jordan R. Green. Automatic speech recognition of conversational speech in individuals with disordered speech. *Journal of Speech, Language, and Hearing Research*, 67(11):4176–4185, 2024.

[14] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.