PRIVACY-ENHANCING FEDERATED CHILD SPEECH RECOGNITION USING MACHINE LEARNING

by

Zihan Lin

May 1 2026

A thesis submitted to the
Faculty of the Graduate School of
the University at Buffalo, State University of New York
in partial fulfilment of the requirements for the
degree of

Master of Science

Department of Computer Science and Engineering

Acknowledgments

I would like to express my sincere gratitude to my advisor, Prof. Wenyao Xu, for his continuous guidance, support, and insightful feedback throughout my research. His expertise and encouragement have been invaluable to the completion of this work.

I am grateful to my colleagues and lab members for their valuable discussions and support, especially Shuwei Hou, Wei Bo, and Chuhui Liu.

Finally, I would like to thank my family for their unwavering support and encouragement.

Abstract

In recent years, Automatic Speech Recognition (ASR) has been extensively studied. However, existing research mainly focuses on optimizing model performance, while receiving little attention to the privacy protection of the data itself, especially the child speech data with explicit features. Since sensitive child speech data is typically siloed across clinics or language pathologists, centralized training is often infeasible. In this paper, we explore a privacy-preserving end-to-end ASR framework that leverages Federated Learning (FL) to fine-tune Whisper under Differential Privacy (DP), aiming to balance recognition performance and data privacy. To mitigate the utility loss introduced by DP, we further introduce FedMem, a local personalization mechanism that enhances performance under privacy constraints by addressing data heterogeneity. Our FL-DP framework reduces the Word Error Rate (WER) from 45.38 for the pre-trained model to 21.32, with FedMem further improving it to 11.88. Experimental results demonstrate that our approach effectively bridges the gap between privacy and utility, incurring only marginal performance degradation (a 2.80 absolute WER increase over the baseline). Furthermore, they also indicate that meaningful differentiation between Typical Development (TD) and Developmental Language Disorder (DLD) samples remains observable under DP constraints, indicating that privacy protection does not weaken the diagnostic effectiveness of clinical indicators.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the development of connected health technologies, clinical assessment and intervention workflows are increasingly extended from traditional clinic settings to home environments, where individuals can perform self-monitoring and interact with mobile applications that support preliminary assessment and intervention under clinical guidance [2, 3]. These connected health technologies transcend the geographical and temporal limitations inherent in traditional periodic clinic visits.

Speech, as a signal closely tied to daily activities, plays an important role in assessing children's health and language development. Children's speech is often characterized by immature speaking patterns, which manifest as speech disfluencies such as repetitions, pauses, and mispronunciation [4]. This challenge is further exacerbated by the pronounced non-identically distributed (Non-IID) nature of children's speech across different developmental stages [5]. Consequently, child ASR models require large amounts of data to achieve robust learning.

To meet this data demand, continuous and large-scale collection of child speech in home environments is a natural and practical solution, as it enables

long-term monitoring, assessment, and intervention workflows. However, in practice, such data are distributed across clinics and households and are constrained by legal restrictions such as CCPA and GDPR [6, 7, 8], as well as data-sharing risks, including potential data leakage caused by network eavesdropping or interception during data transmission [9]. As a result, each client is forced to rely primarily on its own local data for model fine-tuning without access to labeled data from other clients to enrich the model, which can lead to limited coverage of speech variability and reduced generalization to diverse speakers, ages, and audio environments.

In this context, to circumvent the privacy risks and ethical constraints inherent in centralized training, Federated Learning (FL) has emerged as a compelling paradigm for collaborative model training. By keeping data localized and only exchanging model parameters with a central server, FL guarantees that raw speech data is not shared [6, 10]. However, FL alone does not provide formal privacy guarantees, as adversaries can still infer sensitive information from shared model updates [11, 12]. This vulnerability is particularly acute in child ASR scenarios; since neural networks are known to disproportionately memorize unique training patterns [13], the substantial acoustic mismatch in children's voices [14, 15], relative to foundation ASR models primarily pre-trained on adult speech, may leave identifiable feature traces. These 'footprints' can be exploited through inference attacks [16, 11], highlighting the necessity of incorporating additional mechanisms, such as Differential Privacy (DP), into the federated learning framework.

While existing research has begun exploring Federated Learning with Differential Privacy (FL-DP) for foundation ASR models such as Whisper, its application to child speech recognition remains largely unexplored [17]. In this

paper, we present the first empirical benchmark of pediatric ASR under an FL-DP framework, adopting the widely used sample-level Differential Privacy (SL-DP) paradigm. Specifically, we aim to mitigate privacy leakage by injecting noise during training, thereby reducing the influence of child-specific feature traces resulting from the distinct acoustic characteristics of child speech. Meanwhile, to address the challenges posed by data heterogeneity (i.e., non-IID distributions across clients), we further explore FedMem [18], a training-free k-nearest neighbors (kNN) classifier [1], as a personalization mechanism. By leveraging limited local data residing on the client side, FedMem enables client-specific adaptation without compromising privacy. This effectively mitigates the generalization limitations of the FL global model under Non-IID data, thereby reducing performance degradation and partially compensating for the utility loss introduced by DP noise. Our contributions are summarized as follows:

- **System Architecture for Child ASR:** An end-to-end federated Whisper fine-tuning system for child ASR (raw audio remains on-device / on-site).
- **Empirical Analysis of Privacy Trade-offs:** Empirical characterization of privacy-utility-stability tradeoffs in a realistic small-cohort setting, including practical training/compute insights.
- **Validation of clinical utility under DP constraints:** We demonstrate that key diagnostic indicators for differentiating TD and DLD populations in the ENNI dataset remain robust and statistically observable, indicating that privacy-preserving mechanisms do not compromise the model's diagnostic effectiveness [19, 20].
- **Personalization via FedMem Mechanism:** We adapt the FedMem frame-

work to children's ASR to mitigate data heterogeneity (Non-IID), enabling localized client optimization without compromising data privacy.

Chapter 2

System Design & Implementation

This section describes the design and implementation of the proposed FL-DP-FedMem framework with the Whisper model for privacy-preserving child ASR. We first introduce the overall system architecture, followed by the federated training pipeline, the integration of differential privacy mechanisms, and how FedMem personalizes the FL global model for each client.

2.1 System Overview and Local Training Pipeline

In the proposed federated learning (FL) framework, participants — including clinics, hospitals, research institutions, and individual devices — act as clients that locally store and process their own audio data. The raw audio data remains on the local device rather than being transmitted to a central server; only the locally updated models from selected clients are shared with the central server during communication rounds. This design mitigates the risk of exposing sensitive audio, offering robust protection for privacy-sensitive child speech data. For the ASR model, we adopt the Whisper-Small pretrained model, which is

based on a Transformer architecture, and fine-tune it on the Edmonton Narrative Norms Instrument (ENNI) child speech dataset, which contains 300 children with Typical Development (TD) and 77 with developmental language disorder (DLD), using PyTorch. Detailed dataset statistics can be found in [19]. During the preprocessing stage, the dataset is split into training and testing sets using a 70/30 ratio for both the TD and DLD groups, and then evenly distributed across all clients. The 16 kHz audio recordings x are converted into log-Mel spectrograms, and the corresponding reference transcriptions y are tokenized using the Whisper BPE tokenizer to produce token ID sequences. These two components form the input data for the Whisper model. Each local model is trained for 4,000 steps per round with a batch size of 16 audio segments. The learning rate follows a linear schedule, warming up from 0 to 1×10^{-5} over the first 500 steps and then decaying to 0 by the end of training. Finally, the server aggregates the weights of all local models to form a global model, which is then sent back to the clients for the next training round. This process is repeated iteratively across communication rounds. Model performance is evaluated using two metrics, namely loss and Word Error Rate (WER), on the ENNI test dataset. We employ **FedAvg** [10], the standard aggregation algorithm in federated learning, due to its simplicity and robustness. As a foundational framework, FedAvg also provides a consistent baseline that facilitates straightforward extensions to more specialized federated optimization algorithms in future investigations. In our setup, all clients participate in every communication round without client subsampling. Training is conducted for 6 global rounds under synchronous client updates, with fixed random seeds used to ensure reproducibility.

2.2 Privacy layer (SL-DP mechanisms)

Differential Privacy (DP) provides a formal privacy guarantee by ensuring that two neighboring datasets are statistically indistinguishable, even if they differ in a single sample [6]. As a result, an attacker cannot reliably infer whether specific information was included in the training data.

Sample-level differential privacy (SL-DP) aims to protect individual data samples. In this sample level, neighboring datasets differ by only a single data point, rather than by an entire client or device. Before aggregation, each per-sample gradient $g(x_i)$ is clipped based on its ℓ_2 norm with a clipping threshold C to bound the contribution of each individual data point x_i. Specifically, for a batch of size b, the norm of each per-sample gradient $\|g_t(x_i)\|_2$ is constrained to not exceed C.

$$\bar{g}_t(x_i) = \text{Clip}\big(g_t(x_i); C\big) = \frac{g_t(x_i)}{\max\left(1, \frac{\|g_t(x_i)\|_2}{C}\right)} \tag{2.1}$$

After clipping, the batch gradients are summed and independently injected with zero-mean Gaussian noise in accordance with the standard DP-SGD mechanism, thereby preserving an unbiased estimate of the gradient in expectation. To obtain the update for the current batch, this noisy aggregate is subsequently normalized by the batch size b.

$$\tilde{g}_t = \frac{1}{b}\left(\sum_{i\in\mathcal{B}_t} \bar{g}_t(x_i) + \mathcal{N}\left(0, \sigma^2 C^2 \mathbf{I}\right)\right), \tag{2.2}$$

Sample-level differential privacy can be particularly favorable in few-client regimes with relatively large local datasets. By **Concentration Inequalities** [21], as the neighborhood dataset size increases, the aggregated gradient for SL-DP

concentrates within a narrow neighborhood around its expectation, thereby reducing the relative impact of the injected noise. A detailed explanation is provided in Appendix A. This phenomenon is analogous to dispersing ink (fixed amount of DP noise) into the sea (aggregated gradient) rather than into a cup of water, where individual impacts are effectively diluted. Moreover, the zero-mean property of Gaussian noise ensures that the resulting gradient estimate remains unbiased in expectation, helping preserve the overall optimization direction.

2.3 Personalization via FedMem

After federated training, each client receives the fine-tuned global model from the server. On top of this, we incorporate a kNN-based memorization strategy, termed FedMem, which enables client-specific memory retrieval by retrieving semantically closest data from each client's own stored data and adapting the model output accordingly. This local adaptation mechanism effectively addresses the generalization limitations of a unified global model under non-IID conditions, which typically leads to performance degradation. Furthermore, since the datastore is constructed and queried locally, FedMem enables on-device personalization without exposing raw client data to the server. This process is divided into two phases:

Memorization (Datastore Construction): For each client c, a FedMem datastore is constructed using the well-trained global model parameters θ_g together with the client's local dataset. Specifically, for each parallel pair (x, y) in the local training set, where x represents the input audio and y its transcription, we extract the context representations $h_t = f_{\theta_g}(x, y_{<t})$ via a forward pass under teacher forcing. At each time step t, the context representation h_t is paired with

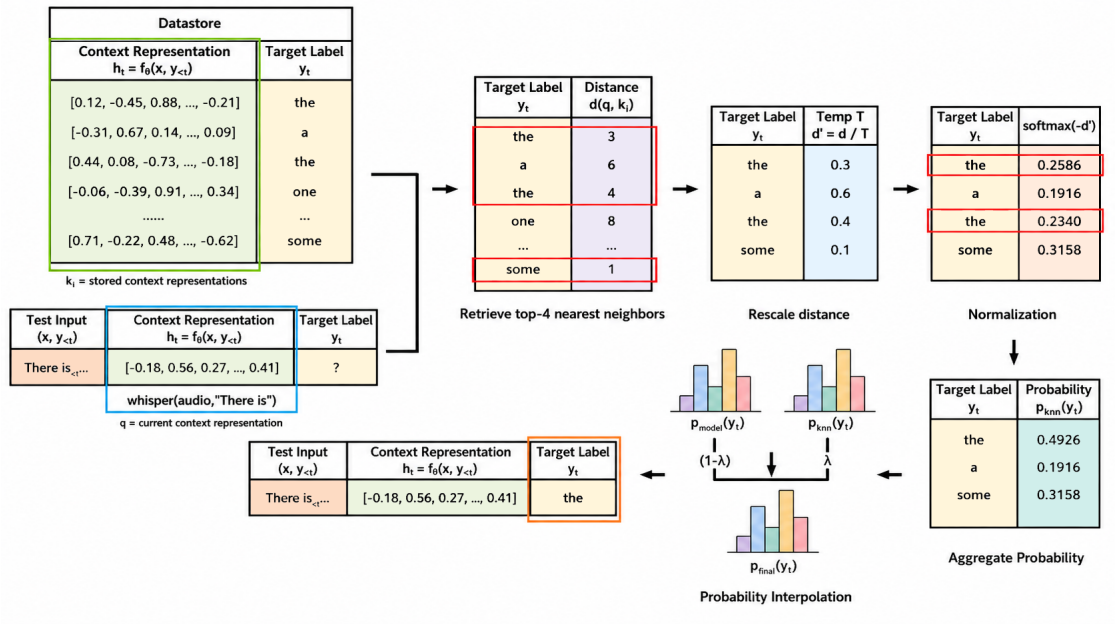

Figure 2.1: Illustration of the FedMem Pipeline. Adapted from [1].

the ground-truth token y_t to form a key-value pair (h_t, y_t). Storing all such pairs over the dataset yields a client-specific datastore, which is used for kNN retrieval during inference.

Retrieval (Inference and Adaptation): During the inference stage, the model performs local adaptation by querying the constructed datastore. At each time step t, the current context representation h_t is used to conduct a kNN search to retrieve the k nearest neighbors from the client-specific datastore via Euclidean (L_2) distance. Smaller distances are assigned higher probabilities, prioritizing the nearest neighbor for next-token prediction. A temperature parameter T is introduced to rescale the distances, thereby smoothing the normalization (softmax) distribution. Subsequently, probabilities associated with identical token IDs among the retrieved neighbors are aggregated to form the final kNN-based distribution p_{kNN} over the target token y_t. The overall FedMem pipeline is illustrated in Figure 2.1.

$$p_{\text{kNN}}(y_t \mid x, \hat{y}_{<t}) \propto \sum_{(h_i, v_i) \in \mathcal{R}} \mathbb{1}_{y_t = v_i} \exp\left(\frac{-d(h_i, f_{\theta_g}(x, \hat{y}_{<t}))}{T}\right) \tag{2.3}$$

The final output distribution is obtained by interpolating the kNN-based distribution p_{kNN} and the probability distribution produced by the fine-tuned model, with an interpolation coefficient $\lambda \in [0, 1]$.

$$p(y_t \mid x, \hat{y}_{<t}) = \lambda\, p_{\text{kNN}}(y_t \mid x, \hat{y}_{<t}) + (1 - \lambda)\, p_{\text{model}}(y_t \mid x, \hat{y}_{<t}) \tag{2.4}$$

Chapter 3

Experimental Setup

3.1 Datasets and evaluation protocol

We fine-tune the Whisper-small model on the ENNI child speech dataset to develop a privacy-preserving child ASR system. To simulate a federated learning environment, the dataset is randomly and equally partitioned between two clients. While this ensures a balanced sample size, the data distribution remains inherently non-IID due to the significant acoustic variability in child speech, such as differences in age, articulation clarity, and disfluency patterns.

This study specifically investigates the interplay between Federated Learning (FL) and Differential Privacy (DP). To quantify the trade-off between privacy guarantees and recognition accuracy under controlled conditions, we adopt a configuration consisting of one central server and two clients. We use the Word Error Rate (WER) as the primary metric to evaluate ASR performance, complemented by training loss to analyze optimization behavior and convergence stability under DP constraints.

3.2 Acoustic Processing and Training Settings

We utilize log-Mel features to represent the acoustic characteristics of child speech, which serve as the input for our ASR model training. As a baseline, Centralized Fine-tuning is conducted for 8,000 optimization steps. This duration is sufficient for convergence, as both the training loss and Word Error Rate (WER) exhibit clear stabilization. Given a batch size of 16, this equates to approximately 128,000 total number of training examples processed ($16 \times 8,000$).

For the Federated Learning (FL) configurations, each client performs one local training phase of 2,000 effective optimizer steps per round across four federated rounds. With a batch size of 16, the process yields approximately 128,000 total number of training examples processed ($2,000 \times 16 \times 4$), directly aligning the total sample exposure and computational effort with the centralized baseline to ensure a fair comparison.

Through a hyperparameter exploration of global rounds and noise multipliers, we observed that four rounds are insufficient for the FL-DP model to fully converge, suggesting that privacy mechanisms demand additional training iterations. Furthermore, while higher noise levels typically hinder performance, a noise multiplier of 1.0 was empirically selected as the optimal configuration to strike an effective balance between model utility and privacy guarantees. These hyperparameter dynamics are analyzed in greater detail in the following section.

3.3 Privacy accounting and DP settings

The privacy budget (ε) is determined by the dataset size, batch size, number of training epochs, and, most critically, the noise multiplier. We compute ε using the Rényi Differential Privacy (RDP) accountant. Following standard practice, we adopt an (ε, δ)-DP formulation with $\delta = 10^{-5}$ and set the clipping norm to 1.0. With all other variables held constant, the privacy budget ε is primarily determined by the noise multiplier σ; therefore, our analysis focuses on the impact of varying noise multipliers on the resulting privacy guarantees.

The sampling rate is approximately 0.008, calculated as the ratio of the batch size (16) to the local dataset size (1,889) per client. Given the 4 federated rounds, the RDP accountant (as implemented in the Opacus library) yields a privacy budget of $\varepsilon = 34.974$ for a noise multiplier of 0.5. In contrast, increasing the noise multiplier to 1.0 significantly strengthens the privacy guarantee by reducing ε to 4.913 and, due to the exponential dependence of the privacy loss on ε, provides substantially more robust protection than the FL-DP baseline configured with a noise multiplier of 0.5.

3.4 Personalization and FedMem settings

To achieve efficient retrieval, FedMem is implemented using FAISS with an IndexIVFPQ configuration. The local datastore of each client is indexed into 2048 coarse clusters (IVF), where each 768-dimensional datastore key is quantized from 3072 bytes to 64 bytes using PQ. At inference time, for each target token y_t, the system identifies the top-64 most relevant clusters via Euclidean (L_2) distance. By searching only within the probed clusters to find k-nearest neighbors, FedMem substantially reduces computational overhead while maintaining robust retrieval performance. Following prior works [1, 18] on k-nearest neighbor retrieval for language tasks, we set $k = 16$ and $T = 10$ as the baseline for our experiments. A grid search over the interpolation coefficient identified $\lambda = 0.3$ as optimal; full details are provided in Appendix B.

Chapter 4

Results

The model, comprising approximately 244M parameters, requires roughly 16 hours for training and 30 minutes for evaluation on a 46 GB GPU. The experimental results are analyzed along two primary dimensions: first, the trade-off between model performance (Word Error Rate) and privacy protection (Privacy Budget ϵ); and second, the preservation of diagnostic utility. Specifically, we evaluate whether the statistical significance between Typical Development (TD) and Developmental Language Disorder (DLD) samples—measured via p-values—is maintained under Differential Privacy (DP) constraints.

4.1 Utility-Privacy Trade-off in Federated Whisper

Since FL and DP typically entail performance degradation, we establish a centralized fine-tuning (CFT) baseline to provide a reference for utility evaluation. From Table 4.1, under identical hyperparameter settings (8,000 optimization steps), the Whisper model reaches a Word Error Rate (WER) of 9.0810. This baseline represents the optimal performance achievable in the absence of privacy-preserving

Experimental Setting	Models	Loss	WER
Baseline (8k steps)	CFT	0.5397	9.0810
	FL	0.5204	9.1702
	FL-DP ($\sigma = 0.5$)	0.5167	23.1360
Local Steps (2k $\rightarrow$ 4k)	FL	0.5685	9.1361
	FL-DP ($\sigma = 0.5$)	0.4304	18.7291
	FL-DP ($\sigma = 1.0$)	0.5230	24.0433
Global Rounds (4 $\rightarrow$ 6)	FL-DP ($\sigma = 0.5$)	0.4059	16.8044
	FL-DP ($\sigma = 1.0$)	0.4853	21.3246

Table 4.1: Eval Performance under Different Hyperparameters.

constraints (i.e., FL and DP).

Building on the centralized baseline, we evaluate the performance under Federated Learning (FL) and Differentially Private FL (DP-FL) configurations. With a conservative hyperparameter setting of 2,000 local optimizer steps per global round across 4 global rounds, the FL-only approach shows a marginal WER increase from 9.0810 to 9.1702, a degradation of approximately 0.1. However, the introduction of Sample-level DP (SL-DP) with a noise multiplier of 0.5 results in a more pronounced WER of 23.1360.

Despite this performance gap, the model maintains effective learning capabilities under privacy constraints, significantly outperforming the pre-trained Whisper model (WER of 45.3841). This resilience is attributed to the moderate local data volume, which facilitates robust gradient concentration. By ensuring that aggregated gradients remain stable, this concentration mitigates the relative impact of injected noise, preventing it from obscuring underlying optimization trends and thus preserving convergence.

To evaluate convergence stability, we increase the local optimization steps per round from 2,000 to 4,000. For standard FL, this adjustment yields a negligible

improvement, with the WER decreasing by only 0.03. In contrast, the DP-FL configuration ($\sigma = 0.5$) shows a substantial gain, with the WER dropping by 4.41.

Recognizing that $\sigma = 0.5$ provides insufficient privacy guarantees, we further explore a more stringent setting with a noise multiplier of 1.0, which aligns with stronger privacy requirements ($\epsilon < 5$). The resulting WER of 24.0433 represents a performance degradation of only 5.31 compared to the $\sigma = 0.5$ setting under identical hyperparameters. Crucially, this modest utility loss enables a near-exponential enhancement in privacy, reducing ϵ from 34.974 to 4.913. This significant decrease in the privacy budget substantially enhances the model's resistance to membership inference and other gradient-based attacks.

Building on these observations, we further examine the impact of global communication rounds. Since increased local training mitigates DP-induced noise, we investigate whether increasing the number of communication rounds further narrows the utility gap or whether the model has already converged. Increasing the number of global rounds from four to six results in reductions in both WER and loss, indicating that the model had not yet converged at round four. However, beyond six rounds, the model appears to approach convergence. Under a noise multiplier of 0.5, the difference in WER between rounds five and six becomes marginal, with a slight increase even observed at round six, suggesting that the model is approaching convergence.

In summary, our experimental results demonstrate that increasing both local optimization steps and global communication rounds improves model utility. As performance begins to stabilize by round 6, we establish the final configuration as 4,000 local steps and 6 global rounds. Regarding privacy settings, a noise multiplier of $\sigma = 1.0$ offers substantially stronger privacy protection than $\sigma = 0.5$.

Experimental Setting	Models	WER
Baseline (8k steps)	CFT	7.6828
	FL	7.7604
	FL-DP ($\sigma = 0.5$)	11.5224
Local Steps (2k $\rightarrow$ 4k)	FL	7.7132
	FL-DP ($\sigma = 0.5$)	10.2849
	FL-DP ($\sigma = 1.0$)	11.6881
Global Rounds (4 $\rightarrow$ 6)	FL-DP ($\sigma = 0.5$)	10.1046
	FL-DP ($\sigma = 1.0$)	11.0822

Table 4.2: Eval Performance after applied FedMem.

Although $\sigma = 1.0$ results in a higher WER (21.3246) compared with $\sigma = 0.5$ (16.8044), the overall performance remains competitive. Therefore, we select $\sigma = 1.0$ as the final noise multiplier, as it yields an optimal trade-off between rigorous privacy guarantees and competitive model utility.

4.2 Improving the Utility–Privacy Trade-off via FedMem

While the preceding results establish the FL-DP baseline configuration, a noticeable utility degradation remains due to noise injection. To mitigate this utility loss, we incorporate the FedMem personalization mechanism under the same experimental setup as the baseline. By addressing data heterogeneity through memorization-based retrieval, FedMem improves model utility while maintaining the same privacy guarantees, thereby improving the utility–privacy trade-off. Table 4.2 shows that FedMem consistently outperforms the baseline in Table 4.1 while maintaining identical privacy guarantees.

As shown in Table 4.2, with $\lambda = 0.3$, the Word Error Rate (WER) for Cen-

tralized Fine-Tuning (CFT) decreases from 9.0810 to 7.6828, while the FL-only approach similarly improves from 9.1702 to 7.7604. Notably, the improvement is more pronounced in the FL-DP setting, suggesting that the personalization mechanism effectively compensates for the utility degradation caused by DP noise. Under the configuration of 6 global rounds and $\sigma = 1.0$, FedMem reduces WER from 21.3246 to 11.0822 with the same privacy guarantees, recovering a substantial portion of the performance lost to DP noise. Moreover, this result narrows the performance gap to only 2.80 relative to the original CFT model in Table 4.1. Overall, these results demonstrate that FedMem effectively mitigates noise-induced degradation and yields a superior privacy–utility trade-off.

4.3 Statistical Significance and Clinical Diagnostic Validity

In this section, we examine whether NTW and NDW exhibit statistically significant differences between the TD and DLD groups. For each ENNI sample, which consists of multiple utterances, total and unique word counts are computed and normalized by the number of utterances, analogous to the computation of Mean Length of Utterance (MLU) [22, 23]. We then compute group-level means and perform statistical significance tests to evaluate whether these between-group linguistic differences remain significant under different training configurations.

This analysis is designed to assess the clinical validity of the privacy-preserving models. Clinical validity is established if: (1) NTW and NDW continue to demonstrate statistically significant differences between the Typical Development (TD) and Developmental Language Disorder (DLD) groups, and

Table 4.3: Significance across Training Settings.

Model	**NTW**			**NDW**		
	TD mean	DLD mean	p-value	TD mean	DLD mean	p-value
Reference Transcript	35.1938	29.2997	0.0663	10.3362	8.7494	0.0664
CFT	35.2438	29.2816	0.0622	10.3307	8.7943	0.0760
FL	35.2547	29.2392	0.0606	10.3065	8.6726	0.0551
FL-DP ($\sigma = 0.5$)	35.9973	36.3402	0.9207	10.1990	8.5983	0.0672
FL-DP ($\sigma = 1.0$)	36.0279	30.3518	0.1057	10.2280	8.6804	0.0743
FL-DP-FM ($\sigma = 1.0$)	35.1323	29.2246	0.0669	10.2653	8.7502	0.0717

(2) the results obtained from Federated Learning with Differential Privacy (FL-DP) models do not exhibit statistically significant deviations from the ENNI gold-standard reference transcripts. Such findings would indicate that FL-DP, despite its stringent privacy guarantees, can produce linguistically faithful transcriptions that preserve clinically meaningful diagnostic markers.

Following the trajectory of our previous hyperparameter exploration, we evaluate a selection of representative models in this experiment to validate their clinical utility. Analysis of the reference transcripts reveal a substantial discrepancy between the TD and DLD groups, with a particularly marked difference in NTW and a 1.6-unit gap in NDW. Since p-values represent the probability of observing the current results under the null hypothesis, the results for both NTW and NDW metrics indicate that, if there were no difference between the TD and DLD groups, the probability of obtaining these specific NTW and NDW results would be only 6.6%. Therefore, TD and DLD show statistically significant differences in this analysis, suggesting a 93.4% likelihood that the NTW and NDW values for the TD group are higher than those for the DLD group. Using the reference transcripts as a baseline, we aim to bring the fine-tuned model's outputs closer to the ENNI gold-standard, thereby effectively preserving these critical clinical markers.

As shown in Table 4.3, both the Centralized Fine-Tuning (CFT) and standard FL models preserve statistically significant differences between the TD and DLD groups. Moreover, their mean NTW and NDW estimates closely align with those of the reference transcript. These findings indicate that both CFT and FL satisfy our predefined criteria for clinical validity, effectively maintaining the diagnostically meaningful linguistic distinctions inherent in the ENNI dataset.

Following the introduction of DP noise, the group means of NTW for TD and DLD exhibited noticeable fluctuations and deviated from those derived from the reference transcripts, whereas NDW remained comparatively stable. In detail, the injected noise caused the ASR model to generate redundant filler tokens, most prominently repeated instances of ”and”. Such repetitions inflate NTW without increasing NDW, thereby weakening the statistical significance of NTW and leading to an increased p-value. As these redundant filler tokens occur stochastically, the NTW results for DLD under a noise multiplier of 0.5 appear to be strongly affected by these random filler tokens during transcription, yielding a relatively high p-value. Conversely, the results at a noise multiplier of 1.0 are free from such random repetitions, resulting in a substantially lower p-value. Crucially, at this higher noise level, the separation between TD and DLD remains robust; the mean differences of approximately 5.6 for NTW and 1.5 for NDW are close to the gold-standard and remain statistically significant. In addition, integrating FedMem under the final configuration (6 global rounds, $\sigma = 1.0$) further bridges the gap to the reference transcript baseline. Specifically, the p-value for NTW is 0.0669, representing only a marginal increase of 0.0006 compared to the reference, while the difference in NDW p-values is similarly small at 0.0053. These results suggest that clinically meaningful linguistic distinctions are preserved even under stronger privacy protection, particularly when

personalization with FedMem is applied.

Chapter 5

Conclusion

In this paper, we presented a privacy-enhancing framework for child automatic speech recognition (ASR). This framework integrates Federated Learning (FL) and Differential Privacy (DP) with the Whisper foundation model, while incorporating FedMem for local personalization. Our results demonstrate that the FL-DP-FedMem framework effectively maintains strong ASR performance under strict privacy constraints. By ensuring that raw speech data remains local and mitigating inference attacks on the global model, this approach is particularly well-suited for privacy-sensitive child ASR scenarios. Furthermore, the framework preserves clinically meaningful diagnostic markers that closely align with those derived from the ENNI gold-standard reference transcripts. Concretely, while operating under robust privacy guarantees (e.g., $\sigma = 1.0$), the framework reduces the WER from 21.3246 to 11.0822 via personalization, and preserves clinically meaningful linguistic distinctions (i.e., differences in NTW and NDW) by showing only minimal deviation from the reference transcripts. [p-value = 0.0669 (NTW) & 0.0717 (NDW)]. Our work demonstrates an improved privacy–utility trade-off, providing a practical foundation for deploying DP-enabled federated

child ASR systems in connected health settings.

For future work, we plan to further explore client-level Differential Privacy (CL-DP) in large-scale settings. As the number of participating clients increases over time, CL-DP is expected to become more effective and stable. In addition, we plan to further enhance model performance by investigating alternative aggregation algorithms. Specifically, we aim to explore frameworks such as FedProx to provide better regularization within Federated Learning, thereby mitigating the challenges posed by data scarcity, data imbalance ($N_{\text{TD}} > N_{\text{DLD}}$), and speaker heterogeneity (Non-IID data).

Appendix A

Concentration Inequalities

The probability:

$$\Pr\left(|X - \mathbb{E}[X]| \geq \varepsilon\right) \tag{A.1}$$

quantifies the likelihood that a random variable X deviates from its expectation E[X] by at least a prescribed threshold ε. The smaller probability indicates X is more concentrated around E[X].

For a finite dataset of size n, Hoeffding's inequality states that

$$\Pr\left(|\bar{X} - \mathbb{E}[\bar{X}]| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) \tag{A.2}$$

This bound shows that, as the dataset size n increases, the probability of observing a large deviation from the expectation decreases exponentially, implying that the sample mean X concentrates within a narrow interval (high-probability deviation range) around the true expectation E[X].

For example, when averaging the outcomes of a small number of dice rolls, the sample mean X may vary widely. In contrast, as the number of rolls increases

to hundreds or thousands, the sample mean X concentrates from an uncertain value to a narrow range around 3.4–3.6 with high probability. Since the sample average stabilizes around 3.4–3.6 for a large number of dataset (rolls time), the addition of DP noise does not substantially alter the overall trend.

Appendix B

FedMem Hyperparameter Selection

In this section, we present the steps for selecting the optimal interpolation coefficient $\lambda \in \{0.1, 0.2, \ldots, 0.9\}$ on the ENNI test set. First, we use CFT with FedMem as the baseline to explore the performance changes after applying different hyperparameters on the test dataset. Under $k = 16$ nearest neighbors and temperature $T = 10$, the results are shown in Table B.1 below:

Table B.1: The ASR performance of CFT-FedMem with different λ

λ	0.1	0.2	**0.3**	0.4	0.5	0.6	0.7	0.8	0.9
WER	7.6634	7.6585	**7.6828**	7.8221	8.5658	8.5691	8.5723	8.5739	8.5772

From the above table, we consider that $[0.3, 0.7]$ is a good range for the interpolation coefficient λ, because the remaining λ values do not show significant performance decay or improvement. Additionally, we want both fine-tuned model and FedMem to contribute a reasonable proportion to the final probability.

Next, we perform inference within the $[0.3, 0.7]$ range using the fine-tuned FL-DP model, and the results are presented separately in Table B.2 below. With the noise multiplier $\sigma = 1.0$, the table shows that when λ exceeds 0.5, the ASR

Table B.2: The ASR performance of FL-DP-FedMem with different λ

λ	**0.3**	0.4	0.5	0.6	0.7
WER	**11.0822**	11.8761	16.3100	16.3902	16.4107

performance degrades by approximately 4.5 in WER. For $\lambda = 0.3$ and 0.4, since we prefer a higher contribution from the fine-tuned model—which is the primary focus of this experiment—and $\lambda = 0.3$ yields better ASR performance, the final choice for λ is 0.3.

Appendix C

Generative AI Use Disclosure

Generative AI tools (e.g., ChatGPT) were used solely for language editing, grammatical polishing, and minor visual enhancement of schematic figures. All technical content, experimental design, and analysis were developed and verified by the authors.

Bibliography

[1] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation, 2021.

[2] Brian M. Caulfield and Steven C. Donnelly. What is connected health and why will it change your practice? *QJM: An International Journal of Medicine*, 106(8):703–707, 2013.

[3] Constantinos S. Pattichis and Andreas S. Panayides. Connected health. *Frontiers in Digital Health*, 1:1, 2019.

[4] Elin T. Thordardottir and Susan Ellis Weismer. Content mazes and filled pauses in narrative language samples of children with specific language impairment. *Brain and Cognition*, 48(2-3):587–592, 2002.

[5] Prashanth Gurunath Shivakumar and Panayiotis Georgiou. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations, 2018.

[6] Jie Fu, Yuan Hong, Xinpeng Ling, Leixia Wang, Xun Ran, Zhiyu Sun, Wendy Hui Wang, Zhili Chen, and Yang Cao. Differentially private federated learning: A systematic review, 2025.

[7] K. J. Mathews and C. M. Bowman. The california consumer privacy act of 2018, 2018. California Civil Code §1798.100–1798.199.

[8] Rachel Cummings and Deven Desai. The role of differential privacy in gdpr compliance. In *In FAT'18: Proceedings of the Conference on Fairness, Accountability, and Transparency. 20.*, 2018.

[9] Mauro Conti, Nicola Dragoni, and Viktor Lesyk. A survey of man in the middle attacks. *IEEE Communications Surveys & Tutorials*, 18(3):2027–2051, 2016.

[10] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2023.

[11] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 739–753, 2019.

[12] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning, 2018.

[13] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks, 2019.

[14] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468, 1999.

[15] Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos. A review of asr technologies for children's speech. *Proceedings of the 2nd Workshop on Child, Computer and Interaction, WOCCI '09*, 11 2009.

[16] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[17] Martin Pelikan, Sheikh Shams Azam, Vitaly Feldman, Jan "Honza" Silovsky, Kunal Talwar, Christopher G. Brinton, and Tatiana Likhomanenko. Enabling differentially private federated learning for speech recognition: Benchmarks, adaptive optimizers and gradient clipping, 2025.

[18] Yichao Du, Zhirui Zhang, Linan Yue, Xu Huang, Yuqing Zhang, Tong Xu, Linli Xu, and Enhong Chen. Communication-efficient personalized federated learning for speech-to-text tasks, 2025.

[19] Phyllis Schneider, Denyse Hayward, and Rita Dubé. Storytelling from pictures using the edmonton narrative norms instrument. *Journal of Speech-Language Pathology and Audiology*, 30:224–238, 12 2006.

[20] Andréanne Gagné and M. Crago. The use of the enni to assess story grammar competency of school-aged french speaking children with and without specific language impairment. *Canadian Journal of Speech-Language Pathology and Audiology*, 34:231–245, 12 2010.

[21] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9, 01 2013.

[22] Roger Brown. *A First Language: The Early Stages*. Harvard University Press, Cambridge, MA, 1973.

[23] Mabel L. Rice, Filip Smolik, Daniel Perpich, Tracy Thompson, Nancy Rytting, and Melissa Blossom. Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, 53(2):333–349, 2010.