

NRC Publications Archive Archives des publications du CNRC

Regularized federated learning for privacy-preserving dysarthric and elderly speech recognition

Zhong, Tao; Geng, Mengzhe; Hu, Shujie; Li, Guinan; Liu, Xunying

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

For the publisher's version, please access the DOI link below. / Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

Publisher's version / Version de l'éditeur:

<https://doi.org/10.21437/Interspeech.2025-1152>

Interspeech 2025, pp. 2103-2107, 2025-08-17

NRC Publications Archive Record / Notice des Archives des publications du CNRC :

<https://nrc-publications.canada.ca/eng/view/object/?id=72bc9373-cbfb-49b7-95e1-5fed571836cb>

<https://publications-cnrc.canada.ca/fra/voir/objet/?id=72bc9373-cbfb-49b7-95e1-5fed571836cb>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at

<https://nrc-publications.canada.ca/eng/copyright>

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site

<https://publications-cnrc.canada.ca/fra/droits>

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at

PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.



Regularized Federated Learning for Privacy-Preserving Dysarthric and Elderly Speech Recognition

Tao Zhong^{1,*}, Mengzhe Geng^{2,*}, Shujie Hu¹, Guinan Li¹, Xunying Liu¹

¹The Chinese University of Hong Kong, China

²National Research Council Canada, Canada

tzhong@se.cuhk.edu.hk, Mengzhe.Geng@nrc-cnrc.gc.ca, sjhu@se.cuhk.edu.hk,
gnli@se.cuhk.edu.hk, xyliu@se.cuhk.edu.hk

Abstract

Accurate recognition of dysarthric and elderly speech remains challenging to date. While privacy concerns have driven a shift from centralized approaches to federated learning (FL) to ensure data confidentiality, this further exacerbates the challenges of data scarcity, imbalanced data distribution and speaker heterogeneity. To this end, this paper conducts a systematic investigation of regularized FL techniques for privacy-preserving dysarthric and elderly speech recognition, addressing different levels of the FL process by 1) parameter-based, 2) embedding-based and 3) novel loss-based regularization. Experiments on the benchmark UASpeech dysarthric and DementiaBank Pitt elderly speech corpora suggest that regularized FL systems consistently outperform the baseline FedAvg system by statistically significant WER reductions of up to 0.55% absolute (2.13% relative). Further increasing communication frequency to one exchange per batch approaches centralized training performance.

Index Terms: speech recognition, dysarthric speech, elderly speech, federated learning, regularization

1. Introduction

While automatic speech recognition (ASR) technologies targeting normal speech have advanced rapidly over the past decades [1, 2], accurate speech recognition in the healthcare domain, particularly for dysarthric and elderly speakers, remains highly challenging to date [3–13]. Dysarthric and elderly speech introduces fundamental challenges for current deep learning-based ASR systems primarily targeting non-aged healthy users, including: **1) substantial mismatch** between such data and normal voices; **2) data scarcity** due to the difficulty of data collection, often limited by speakers' mobility issues [4, 9]; and **3) large speaker heterogeneity** compounding accent, gender, and speech impairments or aging-induced neurocognitive decline [14]. Given the high priority of privacy in healthcare, there has been a growing shift towards decentralized training over centralized approaches, as it mitigates the risk of privacy leakage by limiting the exposure of sensitive data [15].

Federated learning (FL) [16] proves to be an effective method for addressing data privacy concerns, as it enables collaborative model training across decentralized datasets without exchanging raw data. In recent years, there has been increasing attention on applying FL to speech-related tasks for the normal, healthy population, such as keyword spotting [17, 18], speaker verification [19, 20], speech emotion recognition [21, 22], and automatic speech recognition [23–33]. Most prior work in FL-based ASR exclusively targets healthy adult speakers and adopts the Federated Averaging (FedAvg) [16] strategy to ag-

gregate locally trained client models, where the parameters of the individual client are weighted by the proportion of its data samples relative to the total training samples. To constrain local clients from deviating too far from the global model, Federated Averaging with Diversity Scaling (FedAvg-DS) is proposed in [27], while FedProx [34] is investigated in [31] where a proximal term is added to the local training loss. In contrast, so far limited research [12, 35–37] has focused on speech-related FL techniques targeting dysarthric and elderly speakers, with most efforts directed toward Alzheimer's disease (AD) detection and very little toward speech recognition [12].

Current FedAvg-based FL approaches face the following challenges when applied to dysarthric and elderly speech recognition: **1) data scarcity** further exacerbated by splitting the already limited dataset across clients; **2) data imbalance** where speakers with severe speech impairments or linguistic degradation have noticeably fewer words compared to those with milder impairments within a single client; and **3) speaker heterogeneity** among clients. Addressing these challenges necessitates the usage of effective, regularized federated learning techniques.

To this end, this paper conducts a comprehensive exploration of regularized federated learning techniques to address the aforementioned challenges in advancing privacy-preserving dysarthric and elderly speech recognition. Specifically, we investigate regularization at different levels of the FL training process, including: **1) parameter-based regularization**, where a regularization term is added to the local training loss to align the parameters of the local model with those of the global model [34]; **2) embedding-based regularization**, which, inspired by [38, 39], utilizes a regularization term to align intermediate embeddings of the local and global models during local training; and **3) novel loss-based regularization**, where local intermediate embeddings are passed through a frozen global model to generate pseudo-logits. These pseudo-logits are then aligned with local predictions using Kullback-Leibler (KL) divergence [40], which is further incorporated into the local training loss. Performance evaluation is conducted on two benchmark healthcare datasets: **1) UASpeech** [41] dysarthric speech corpus; and **2) DementiaBank Pitt** [42] for elderly speech, with FedAvg [16] serving as the base model aggregation strategy. In addition, the effects of applying regularization techniques at multiple positions, combining different regularization methods, and varying communication frequency are further analyzed.

The main contributions of this paper are as follows:

1) To the best of our knowledge, this paper presents the first systematic investigation of regularized FL techniques for privacy-preserving speech recognition in healthcare, with a focus on dysarthric and elderly populations. Different elements of the FL training process are explored, including parameter-based, embedding-based and the novel loss-based regulariza-

*These authors contributed equally.

tion. In contrast, prior research on regularized federated learning in speech recognition has been notably limited [27, 31], and solutions addressing the challenges of FL-based dysarthric and elderly speech recognition have been rarely visited [12].

2) Experiments on the benchmark UASpeech dysarthric speech and DementiaBank Pitt elderly speech corpora suggest that statistically significant word error rate (WER) reductions of up to 0.54% absolute (1.69% relative) and 0.55% absolute (2.13% relative) can be respectively obtained over the baseline FedAvg system without regularization. Combining the regularization techniques leads to further performance improvements.

The rest of the paper is organized as follows. Section 2 describes the standard FL based ASR systems. The three regularization techniques, i.e., parameter-, embedding- and loss-based regularization, are detailed in Section 3. Section 4 presents experiments and analysis on UASpeech and DementiaBank Pitt. Section 5 concludes the paper and discusses future work.

2. Federated Learning Based ASR

Unlike traditional distributed training [43] which first aggregates data from multiple sources and then redistributes the data among learners, federated learning (FL) preserves data locality to comply with privacy constraints. However, such a design inherently results in heterogeneous data distributions across clients. As illustrated in Fig. 1, each client operates a local server for data storage and computation, interacting solely with a central server. During each communication round, only model parameters or gradients are exchanged between clients and the server. Local updates are sent to the central server, where they are aggregated into a global model (Fig. 1(a)). The global model is then redistributed to clients as the starting point for client-side computation in the next round (Fig. 1(b), dashed line).

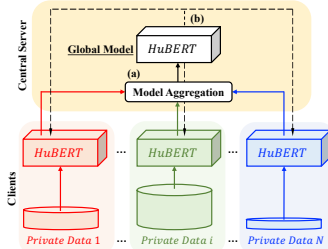


Figure 1: Illustration of federated learning based HuBERT [44] ASR system. Following FedAvg [16], each communication round involves: (a) aggregating the parameters of locally trained client-side models into a global model using a weighted sum, and (b) redistributing the global model to clients.

The widely used aggregation strategy for FL-based ASR is FedAvg [16], which computes global parameters as a weighted sum of client parameters, with weights proportional to each client’s data fraction. Given the communication overhead [33], synchronization is performed at carefully chosen frequency¹.

3. Regularized Federated Learning

Federated learning inherently introduces data heterogeneity among clients, which is further compounded by the challenges of data scarcity, data imbalance, and speaker diversity widely observed in dysarthric and elderly speech. To mitigate this, three regularization techniques at different levels are investi-

¹The impact of communication frequency is analyzed in Section 4.

gated, including parameter-based, embedding-based, and loss-based regularization. For model aggregation, we adopt FedAvg [16] as the default strategy.

3.1. Parameter-based regularization

Following FedProx [34], the global model parameters from the previous communication round are taken as a reference to regularize the local training process of the clients. As shown in the upper right of Fig. 2, the squared L2 norm \mathcal{R}_{para} , which measures the Euclidean distance between the reference global model and the local model, is incorporated into the local training loss as a regularization term, given as:

$$\mathcal{R}_{para} = \|\mathbf{W}_i - \overline{\mathbf{W}}\|_2^2 \quad (1)$$

where \mathbf{W}_i represents the local model parameters of the i^{th} client at the current round, while $\overline{\mathbf{W}}$ denotes the global model parameters from the previous round.

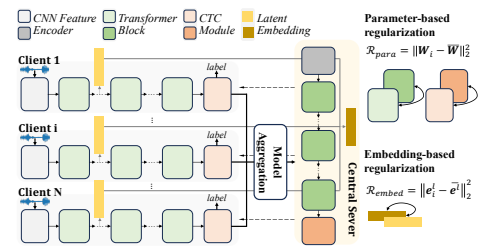


Figure 2: Schematic of (a) parameter-based regularization (upper right) and (b) embedding-based regularization (lower right) applied to the FedAvg-based HuBERT ASR system. Components with darker colors are on the central server side, while lighter-colored components are on the local client side.

3.2. Embedding-based regularization

Building on [38,39], during each communication round, we extract latent embeddings from the local models, compute the average embedding over all utterances for each client, and then aggregate these averaged embeddings from all clients on the central server using a weighted sum². As demonstrated in the lower right of Fig. 2, the aggregated embedding serves as a reference to constrain the local models’ latent embeddings in the next round, where the local training loss integrates the squared L2 norm between these embeddings. This regularization term \mathcal{R}_{embed} is given as:

$$\mathcal{R}_{embed} = \|e_i^l - \overline{e}^l\|_2^2 \quad (2)$$

where e_i^l represents the embedding of the i^{th} client after the l^{th} Transformer block, and \overline{e}^l denotes the aggregated embedding.

3.3. Loss-based regularization

To further investigate regularization across different aspects of the training process, we propose a novel loss-based regularization approach. As illustrated in Fig. 3, rather than directly constraining the latent embedding, we feed the local embedding of the i^{th} client into the global model from the previous communication round and execute a feedforward pass to derive the output probability distribution (Fig. 3, red bold line). This distribution is then leveraged to regularize the output distribution of the i^{th} client by incorporating the Kullback-Leibler

²The weight is the client’s proportion of the total training data.

(KL) divergence [40] between these two distributions into the local training loss. The regularization term \mathcal{R}_{loss} is defined as:

$$\mathcal{R}_{loss} = D_{KL}(\hat{\mathbf{y}}_i \parallel \tilde{\mathbf{y}}_i^l) \quad (3)$$

where $D_{KL}(\cdot)$ denotes the KL divergence. $\hat{\mathbf{y}}_i$ represents the output distribution of the i^{th} client, and $\tilde{\mathbf{y}}_i^l$ denotes the output distribution obtained by passing the i^{th} client’s embedding after the l^{th} Transformer block through the global model.

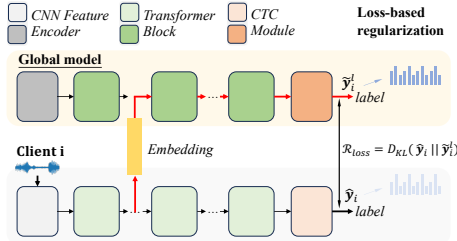


Figure 3: Schematic of the proposed loss-based regularization applied to the FedAvg-based HuBERT ASR system. For simplicity, one client is shown as an example. Darker-colored components represent the global model, while lighter-colored components correspond to the local model.

The embedding-based regularization and loss-based regularization can be applied at multiple positions in the model. Furthermore, as these three regularization techniques operate at different levels, their usage can be combined.

4. Experiments and Results

4.1. Task description

The English UASpeech corpus [41] is the largest publicly available and widely used dataset for dysarthric speech recognition. It comprises an isolated word recognition task with approximately 103 hours of speech data from 29 speakers, among whom 16 are dysarthric speakers and 13 are healthy control speakers. The dataset includes 155 common words and 300 uncommon words and is further divided into three blocks B1, B2 and B3. The same 155 common words are used across all blocks, while the 300 uncommon words differ between blocks. In our experiments, we focus **exclusively on the 16 dysarthric speakers** and exclude the healthy control speakers. B1 and B3 are used as for training, while B2 is used for evaluation. After removing silence, the training set and the test set contain 17.8 hours (52785 utterances) and 9 hours of audio (26520 utterances) in total, respectively. Speech intelligibility assessment is available for the 16 dysarthric speakers, divided into four groups: “very low” (VL), “low” (L), “mid” (M) and “high” (H). **The English DementiaBank Pitt** [42] corpus is the most widely used publicly available dataset for speech-based diagnosis of Alzheimer’s Disease (AD). It comprises 33 hours of cognitive impairment assessment interviews between 292 elderly participants and the associated clinical investigators. The training set includes 688 speakers (244 elderly participants, 444 investigators), while the development and evaluation sets³ respectively consist of 119 (43 elderly, 76 investigators) and 95 speakers (48 elderly, 47 investigators). There is no speaker overlap

³The evaluation set includes the 48 speakers’ Cookie Theft recordings from the ADReSS challenge test set [45], while the development set contains their recordings from other tasks, if available.

between the training set and either the development or evaluation sets. Silence stripping [3] produces a 15.7-hour training set (29682 utterances), a 2.5-hour development set (5103 utterances), and a 0.6-hour evaluation set (928 utterances).

4.2. Experiment setup

Model configuration: We adopt the state-of-the-art HuBERT [44] model⁴ fine-tuned on 960 hours of Librispeech [46] as the foundation for performing federated learning⁵ on dysarthric and elderly speech. The default model aggregation strategy is FedAvg [16]. A Connectionist Temporal Classification (CTC) [47] model with a single fully connected layer is added on top of the CNN feature encoder and a stack of 24 Transformer blocks. In each communication round, the clients perform local training for one epoch, with the number of communication rounds set to 100. The penalty weights for parameter-based, embedding-based, and loss-based regularization are empirically set to 0.01, 0.001, and 0.01, respectively. Parameter-based regularization is applied to all model parameters except those of the CNN feature encoder, while embedding-based and loss-based regularization can be applied after the 6th, 12th, 18th, or 24th Transformer block, or in any combination of these positions. Experiments are conducted using two Nvidia A40 GPUs. A matched pairs sentence-segment word error (MAPSSWE) [48] based statistical significance test is performed with a significance level of $\alpha = 0.05$.

Client partitioning: For the UASpeech dysarthric speech task, the training data of each of the 16 dysarthric speakers is assigned to a separate client, resulting in a total of 16 clients, each containing 0.71 to 1.54 hours of audio (1785 to 3570 utterances). For the DementiaBank Pitt elderly speech task, all conversations between each participant in the training set and the corresponding investigators are extracted, producing 244 conversation pairs in total. These 244 pairs are randomly divided into 10 non-overlapping sets, with each set assigned to one client, resulting in 10 clients with 1.35 to 1.75 hours of audio (2545 to 3437 utterances) per client.

4.3. Result analysis

Experiments on dysarthric speech: Table 1 compares parameter-based, embedding-based and loss-based regularization for federated learning on UASpeech. Several trends can be observed: **1)** All three regularization techniques lead to performance improvements over the baseline FedAvg system without regularization (Sys.2-14 vs. Sys.1), while the proposed loss-based regularization produces better performances (Sys.9-14 vs. Sys.2-8). **2)** By applying embedding-based (Sys.7-8) or loss-based regularization (Sys.13-14) at different positions⁶, with statistically significant overall WER reductions of up to 0.38% abs. (1.19% rel., Sys.8 vs. Sys.1) and 0.54% abs. (1.69% rel., Sys. 14 vs. Sys.1) over the baseline FedAvg system, respectively. **3)** Combining the three regularizations leads to further performance improvement⁷(Sys.15). **4)** Compared with centralized training (Sys.0), there remains a performance gap in federated learning, highlighting the challenges of FL-based dysarthric speech recognition due to its nature.

Experiments on elderly speech: The performance of the three regularization techniques on DementiaBank Pitt is presented in

⁴<https://huggingface.co/facebook/hubert-large-ls960-ft>

⁵<https://apple.github.io/pfl-research/>

⁶Sys.7 combines the two best-performing positions among Sys.3-6, whereas Sys.13 integrates the two best positions from Sys.9-12.

⁷Weight is set as 0.1, 0.1, 1 for para., embed. and loss regularization.

Table 1: Performance of parameter (“para.”), embedding (“embed.”) and loss based regularization for federated learning on UASpeech. Here “6”, “12”, “18” and “24” refer to the 6th, 12th, 18th, or 24th Transformer block. “VL”, “L”, “M” and “H” are “very low”, “low”, “mid” and “high” speech intelligibility. † denote a statistically significant improvement ($\alpha = 0.05$) obtained over the baseline FedAvg system (Sys.1).

Sys.	Regularization				UASpeech WER%						
	Method	Position				Speech Intelligibility				All	
		6	12	18	24	VL	L	M	H		
0	centralized training				64.03	34.89	21.37	6.05	28.87		
1	✗				72.39	38.78	22.78	6.71	32.04		
2	para.	-			72.22	38.56	22.67	6.44 [†]	31.83		
3	embed.	✓			72.25	38.57	22.66	6.46 [†]	31.85		
4			✓		72.23	38.55	22.67	6.43 [†]	31.83		
5				✓	72.20	38.58	22.69	6.42 [†]	31.83		
6					72.10	38.58	22.67	6.42 [†]	31.81		
7			✓	✓	71.95	38.50	22.65	6.40 [†]	31.70 [†]		
8		✓	✓	✓	71.74 [†]	38.44 [†]	22.62	6.36 [†]	31.66 [†]		
9		loss	✓			71.80	38.48	22.63	6.38 [†]	31.68 [†]	
10				✓		71.70 [†]	38.44 [†]	22.60	6.32 [†]	31.62 [†]	
11				✓	71.75 [†]	38.44 [†]	22.62	6.35 [†]	31.65 [†]		
12					71.95	38.50	22.65	6.40 [†]	31.75		
13			✓	✓	71.62 [†]	38.30 [†]	22.60	6.25 [†]	31.54 [†]		
14	✓		✓	✓	71.55 [†]	38.22 [†]	22.58	6.20 [†]	31.50 [†]		
15	para.+embed.+loss				71.50[†]	38.22[†]	22.50	6.19 [†]	31.45[†]		

Table 2. Trends similar to those on dysarthric speech (Table 1) are observed, with statistically significant overall WER reductions of up to 0.43% abs. (1.67% rel.) and 0.55% abs. (2.13% rel.) obtained by embedding-based regularization (Sys.8) and loss-based regularization (Sys.14) over the baseline FedAvg system without regularization (Sys.1), respectively.

4.4. Impact of communication frequency

As communication overhead is a major challenge in federated learning [33], we further investigate the impact of communication frequency on model performance. As shown in Fig. 4, we vary the number of local updates performed by clients in each round, including an extreme case where communication occurs after every batch. To ensure a fair comparison, the total number of communication rounds is adjusted to keep the total training steps the same. Communicating after every batch approaches the performance of centralized learning (Fig. 4, dashed line) on both datasets while keeping the data local, but this comes at the cost of 9 times increase in training time compared to the default setting of communicating once per epoch. Furthermore, the proposed loss-based regularization demonstrates a consistent and statistically significant performance improvement over the unregularized FedAvg systems (Fig. 4, grey lines) across all communication frequencies, except in the 1-batch scenario.

On top of the communication costs incurred in the standard FedAvg training, the additional costs for these three regularization methods are: 1) 0 for parameter-based and loss-based embedding, as the former contrasts the global model with the local model while the latter feeds the local embedding to the global model for the loss comparison, both performed on the client side; 2) $O(d_{embed})$ for embedding-based regularization for transmitting the embeddings, where d_{embed} is the size of the embeddings. When all three regularization techniques are

Table 2: Performance of parameter (“para.”), embedding (“embed.”) and loss based regularization for federated learning on DementiaBank Pitt. “Dev” and “Eval” are the development and evaluation sets. “PAR” and “INV” are elderly participants and clinical investigators. † denote a statistically significant improvement ($\alpha = 0.05$) obtained over the baseline FedAvg system (Sys.1). Other naming conventions follow Table 1.

Sys.	Regularization				DementiaBank Pitt WER%						
	Method	Position				Dev		Eval		All	
		6	12	18	24	PAR	INV	PAR	INV		
0	centralized training				31.62	16.43	22.93	15.87	23.52		
1	✗				34.14	18.52	25.38	16.76	25.80		
2	para.	-			33.84	18.28	25.15	16.30	25.54		
3	embed.	✓			33.90	18.33	25.16	16.30	25.58		
4			✓		33.86	18.26	25.15	16.28	25.54		
5				✓	33.80	18.24	25.14	16.28	25.52		
6					33.82	18.23	25.12	16.25	25.51		
7			✓	✓	33.70 [†]	18.20	25.05	16.25	25.44 [†]		
8		✓	✓	✓	33.65 [†]	18.18	25.01	16.22	25.37 [†]		
9		loss	✓			33.65 [†]	18.23	25.04	16.25	25.44 [†]	
10				✓		33.65 [†]	18.20	25.01	16.22	25.40 [†]	
11				✓	33.68 [†]	18.22	25.01	16.25	25.44 [†]		
12					33.70 [†]	18.25	25.05	16.25	25.50		
13			✓	✓	33.60 [†]	18.10	25.00	16.15	25.32 [†]		
14	✓		✓	✓	33.50 [†]	18.00	24.91	16.11	25.25 [†]		
15	para.+embed.+loss				33.45[†]	18.00	24.85[†]	16.06	25.21[†]		

combined, the additional costs are $O(d_{embed})$.

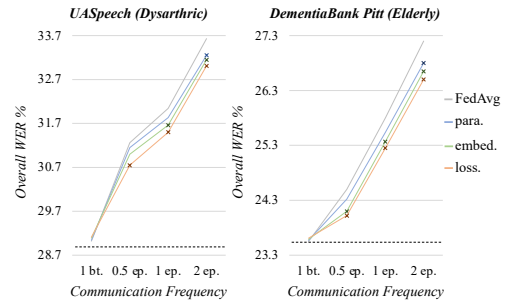


Figure 4: Impact of communication frequency on the performance (WER%) of federated learning. “bt.” and “ep.” stand for batch and epoch. × denotes a stat. significant improvement ($\alpha = 0.05$) over the baseline FedAvg system. The dashed line represents the performance of centralized learning (WER 28.87% for UASpeech and 23.52% for DementiaBank Pitt).

5. Conclusions

This paper systematically investigates regularized FL techniques for privacy-preserving dysarthric and elderly speech recognition, i.e., parameter-, embedding-, and novel loss-based regularizations. Experiments on UASpeech and DementiaBank Pitt show that regularized FL systems consistently outperform FedAvg, while increasing communication frequency narrows the gap to centralized training. Future research will focus on speech-pattern driven regularization techniques.

6. Acknowledgements

This research is supported by Hong Kong RGC GRF grant No. 14200220, 14200021, 14200324 and Innovation Technology Fund grant No. ITS/218/21.

7. References

- [1] L. Dong, S. Xu *et al.*, “Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition,” in *ICASSP*, 2018.
- [2] A. Gulati, J. Qin *et al.*, “Conformer: Convolution-augmented Transformer for Speech Recognition,” in *INTERSPEECH*, 2020.
- [3] Z. Ye, S. Hu *et al.*, “Development of the CUHK Elderly Speech Recognition System for Neurocognitive Disorder Detection Using the Dementiabank Corpus,” in *ICASSP*, 2021.
- [4] S. Liu *et al.*, “Recent Progress in the CUHK Dysarthric Speech Recognition System,” *IEEE/ACM T-ASLP*, 2021.
- [5] T. Wang, J. Deng *et al.*, “Conformer Based Elderly Speech Recognition System for Alzheimer’s Disease Detection,” in *INTERSPEECH*, 2022.
- [6] M. Geng, X. Xie *et al.*, “Speaker Adaptation Using Spectro-Temporal Deep Features for Dysarthric and Elderly Speech Recognition,” *IEEE/ACM T-ASLP*, 2022.
- [7] Z. Yue, E. Loweimi *et al.*, “Acoustic modelling from raw source and filter components for dysarthric speech recognition,” *IEEE/ACM T-ASLP*, 2022.
- [8] M. K. Baskar, T. Herzig *et al.*, “Speaker adaptation for Wav2vec2 based dysarthric ASR,” in *INTERPSEECH*, 2022.
- [9] S. Hu, X. Xie *et al.*, “Self-Supervised ASR Models and Features for Dysarthric and Elderly Speech Recognition,” *IEEE/ACM T-ASLP*, 2024.
- [10] Z. Jin, M. Geng *et al.*, “Personalized Adversarial Data Augmentation for Dysarthric and Elderly Speech Recognition,” *IEEE/ACM T-ASLP*, 2023.
- [11] H. Wang, Z. Jin *et al.*, “Enhancing Pre-trained ASR System Fine-tuning for Dysarthric Speech Recognition using Adversarial Data Augmentation,” in *ICASSP*, 2024.
- [12] W.-T. Hsu, C.-P. Chen *et al.*, “A Cluster-based Personalized Federated Learning Strategy for End-to-End ASR of Dementia Patients,” in *INTERSPEECH*, 2024.
- [13] H. Wang, X. Xie *et al.*, “Phone-purity Guided Discrete Tokens for Dysarthric Speech Recognition,” *arXiv preprint arXiv:2501.04379*, 2025.
- [14] I. Kodrasi and H. Bourlard, “Spectro-temporal sparsity characterization for dysarthric speech detection,” *IEEE/ACM T-ASLP*, 2020.
- [15] S. M. Williamson and V. Prybutok, “Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare,” *Applied Sciences*, 2024.
- [16] B. McMahan, E. Moore *et al.*, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in *AISTATS*, 2017.
- [17] D. Leroy, A. Coucke *et al.*, “Federated learning for keyword spotting,” in *ICASSP*, 2019.
- [18] A. Hard, K. Partridge *et al.*, “Training keyword spotting models on non-iid data with federated learning,” in *INTERSPEECH*, 2020.
- [19] F. Granqvist, M. Seigel *et al.*, “Improving on-device speaker verification using federated learning with privacy,” in *INTERSPEECH*, 2020.
- [20] L. Zhang, L. Liu *et al.*, “Stealthy backdoor attack towards federated automatic speaker verification,” in *ICASSP*, 2024.
- [21] S. Latif, S. Khalifa *et al.*, “Federated learning for speech emotion recognition applications,” in *IPSN*, 2020.
- [22] T. Feng and S. Narayanan, “Semi-fedSER: Semi-supervised learning for speech emotion recognition on federated learning using multiview pseudo-labeling,” in *INTERSPEECH*, 2022.
- [23] D. Dimitriadis, R. G. Ken’ichi Kumatani *et al.*, “A federated approach in training acoustic models,” in *INTERSPEECH*, 2020.
- [24] D. Guliani, F. Beaufays *et al.*, “Training speech recognition models with federated learning: A quality/cost framework,” in *ICASSP*, 2021.
- [25] W. Yu, J. Freiwald *et al.*, “Federated learning in ASR: Not as easy as you think,” in *ITG SpeechCom*, 2021.
- [26] X. Cui, S. Lu *et al.*, “Federated acoustic modeling for automatic speech recognition,” in *ICASSP*, 2021.
- [27] K. Nandury, A. Mohan *et al.*, “Cross-silo federated training in the cloud with diversity scaling and semi-supervised learning,” in *ICASSP*, 2021.
- [28] Y. Gao, T. Parcollet *et al.*, “End-to-end speech recognition from federated acoustic models,” in *ICASSP*, 2022.
- [29] S. S. Azam, T. Likhomanenko *et al.*, “Importance of Smoothness Induced by Optimizers in FL4ASR: Towards Understanding Federated Learning for End-to-End ASR,” in *ASRU*, 2023.
- [30] S. S. Azam, M. Pelikan *et al.*, “Federated Learning for Speech Recognition: Revisiting Current Trends Towards Large-Scale ASR,” in *NeurIPS*, 2023.
- [31] M. Pelikan, S. S. Azam *et al.*, “Federated learning with differential privacy for end-to-end speech recognition,” *arXiv preprint arXiv:2310.00098*, 2023.
- [32] X. Kan, Y. Xiao *et al.*, “Parameter-Efficient Transfer Learning under Federated Learning for Automatic Speech Recognition,” *arXiv preprint arXiv:2408.11873*, 2024.
- [33] Y. Du, Z. Zhang *et al.*, “Communication-Efficient Personalized Federated Learning for Speech-to-Text Tasks,” in *ICASSP*, 2024.
- [34] T. Li, A. K. Sahu *et al.*, “Federated Optimization in Heterogeneous Networks,” in *MLSys*, 2020.
- [35] S. I. A. Meerza, Z. Li *et al.*, “Fair and Privacy-Preserving Alzheimer’s Disease Diagnosis Based on Spontaneous Speech Analysis via Federated Learning,” in *EMBC*, 2022.
- [36] S. T. Arasteh, C. D. Rios-Urrego *et al.*, “Federated learning for secure development of AI models for Parkinson’s disease detection using speech from different languages,” in *INTERSPEECH*, 2023.
- [37] S. Kalabakov, M. Gonzalez-Machorro *et al.*, “A Comparative Analysis of Federated Learning for Speech-Based Cognitive Decline Detection,” in *INTERSPEECH*, 2024.
- [38] Y. L. Tun, C. M. Thwal *et al.*, “Federated learning with intermediate representation regularization,” in *BigComp*, 2023.
- [39] R. Greidi and K. Cohen, “Sparse Training for Federated Learning With Regularized Error Correction,” *JSTSP*, 2024.
- [40] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Ann. Math. Stat.*, 1951.
- [41] H. Kim, M. Hasegawa-Johnson *et al.*, “Dysarthric speech database for universal access research,” in *INTERSPEECH*, 2008.
- [42] J. T. Becker, F. Boiler *et al.*, “The natural history of alzheimer’s disease: description of study cohort and accuracy of diagnosis,” *Arch. Neurol.*, 1994.
- [43] J. Verbraeken, M. Wolting *et al.*, “A survey on distributed machine learning,” *CSUR*, 2020.
- [44] W.-N. Hsu, B. Bolte *et al.*, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” *IEEE/ACM T-ASLP*, 2021.
- [45] S. Luz, F. Haider *et al.*, “Alzheimer’s Dementia Recognition through Spontaneous Speech: The ADReSS Challenge,” in *INTERSPEECH*, 2020.
- [46] V. Panayotov, G. Chen *et al.*, “Librispeech: An ASR corpus based on public domain audio books,” in *ICASSP*, 2015.
- [47] A. Graves, S. Fernández *et al.*, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *ICML*, 2006.
- [48] M. Bisani and H. Ney, “Bootstrap estimates for confidence intervals in asr performance evaluation,” in *ICASSP*, 2004.