Interspeech 2024
1-5 September 2024, Kos, Greece

A Cluster-based Personalized Federated Learning Strategy for End-to-End ASR of Dementia Patients

Wei-Tung Hsu[1], *Chin-Po Chen*[1], *Yun-Shao Lin*[1], *Chi-Chun Lee*[1]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan

wthsu110061542@gapp.nthu.edu.tw, Chen.JackCP@inventec.com, yunshaolin@gmail.com, cclee@ee.nthu.edu.tw

Abstract

Automatic speech recognition (ASR) is crucial for all users, but adapting it for Alzheimer's disease (AD) faces challenges due to irregular speech patterns and privacy concerns. Federated learning (FL), a privacy-preserving algorithm, is a solution. However, FL ASR suffers from acoustic and text heterogeneities. While advanced model-based and cluster-based FL methods aim to address the issue, they lack a direct mechanism for high intra-speaker heterogeneity exhibited by AD individuals and ASR-related properties. This study presents cluster-based personalized federated learning (CPFL), a strategy mitigating heterogeneity by clustering ASR output token using the proposed CharDiv, a metric for pause and word usage distributions. Evaluation on the ADReSS challenge dataset shows a 3.6% improvement in word error rate (WER). Analysis of per-cluster WER improvements and CharDiv distributions indicates reduced heterogeneity, emphasizing pause usage as a potential key factor in AD-oriented ASR.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Automatic speech recognition (ASR) is vital in modern life, enhancing human-machine interaction and showing potential in assisting early detection of neurocognitive disorders like Alzheimer's disease (AD) [1, 2]. AD, the most common type of dementia that is irreversible and incurable, mostly affecting the elderly [3], poses two key challenges for effective ASR generalization. First, distinctive speaking traits of individuals with AD, like disfluent speeches with longer pauses [4], create great intra-speaker variations, complicating the task compared to regular speech patterns [5]. Second, privacy preservation, crucial for those with diseases concerned about personal information exposure, adds complexity to the challenge, limiting data usage in clinical facilities. These issues highlight the need for large-scale data, hindered by privacy concerns. Federated learning (FL), an algorithm offering a non-data-sharing training scheme that encourages joint training from clinical facilities [6], is an appealing approach for privacy-preserved AD-oriented ASR.

FL for ASR holds promise but presents complexities beyond other FL applications [7, 8], which are compounded by added acoustic and text heterogeneity, involving diverse vocalizations, prosody, and word usage in individuals with AD. Recent FL approaches tackle heterogeneity in two main aspects: model and cluster. For the model aspect, approaches fall into two categories: generalization and personalization. Generalization algorithms, like FedAvg-DS [9] and FedProx [10], regulate individual client weights before aggregation to achieve global models. However, a one-size-fits-all model may not suit highly heterogeneous datasets. Personalization algorithms, such as federated mutual learning (FML) [11], implement dual-model approaches for each client, facilitating the information exchange model and the refining model for local data. While these methods provide corresponding models for different clients, locally personalized models may struggle with significant heterogeneity within client data, where differences within samples from the same client can exceed those between samples from different clients, leading to reduced performance. In contrast, FL approaches for the cluster aspect enable similar samples or clients to have corresponding cluster models targeting specific problems, offering a more effective solution to the challenges in AD-oriented ASR.

While recent cluster-based FL methods show promise in addressing heterogeneity, their suitability for AD-oriented ASR remains unclear, particularly concerning the clustering unit and metric. For instance, personalized clustered FL (PCFL) [12] clusters clients based on model weights. However, in the context of AD, where intra-speaker heterogeneity can exceed inter-speaker heterogeneity, a more rational approach might involve clustering based on individual samples rather than grouping clients or speakers. Similarly, community-based FL (CBFL) [13] clusters clients' data using embeddings derived from the data. However, for ASR end-to-end models, a more suitable clustering approach involves grouping based on ASR output for accurate result prediction. These gaps underscore the urgent need for effective cluster-based FL strategies to tackle the heterogeneity specific to AD-oriented ASR tasks, considering both the appropriate clustering unit and metric, to optimize FL-based ASR performance.

This study introduces a novel cluster-based FL method tailored for AD-oriented ASR training, with a focus on mitigating heterogeneity issue through text token-based clustering strategy. Our approach comprises two key components: the cluster-based ASR system and the clustering metric. The cluster-based ASR system utilizes the cluster-based personalized FL (CPFL) strategy, which organizes clusters and assigns clients to train ASR models federally, based on data from corresponding clusters. Clusters are formed using clients' samples with similar character diversity (CharDiv), our proposed clustering metric capturing pause and word usage distributions based on ASR's output tokens while ensuring privacy by not revealing actual spoken content. This helps mitigate heterogeneity of FL learning strategy. Evaluation on the ADReSS challenge dataset [14] including samples from healthy elderly individuals and those with AD, demonstrates our approach's superiority over various other FL approaches, achieving a 3.6% reduction in word error rate (WER) [15]. Analysis of per-cluster WER improvements and CharDiv distributions provides supporting evidence for the effectiveness of our method.

10.21437/Interspeech.2024-470

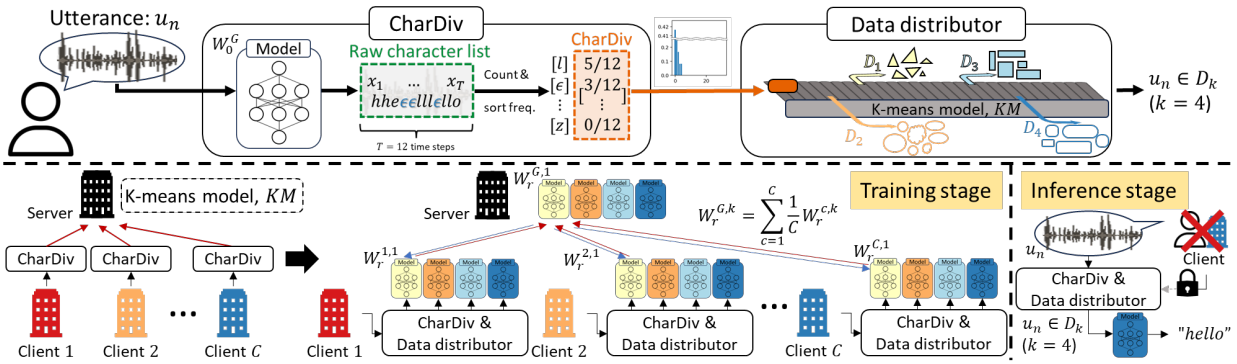

Figure 1: *The cluster-based personalized federated learning (CPFL) strategy groups samples with similar character diversity (CharDiv) into clusters, where clients federally train these samples to create a model for decoding others within the same cluster.*

2. Methodology

This study proposes the use of character diversity (CharDiv) as a clustering metric, derived from ASR's tokens, for clustering-based FL [12, 13]. Section 2.1 presents the dataset, and section 2.2 outlines the proposed CharDiv-clustered FL strategy. Section 2.2.1 illustrates the cluster-based personalized FL (CPFL) framework along with the K-means model training for clustering, and explains the use of trained models at inference. Section 2.2.2 describes the computation of CharDiv. CharDiv-clustered FL is depicted in Figure 1, with code available on GitHub [1].

2.1. Dataset

ADReSS challenge dataset [14], provided by DementiaBank [16], serves as a benchmark for predicting dementia patients using spontaneous speech, with a balanced distribution of age, gender, and health conditions. Comprising transcribed speech recordings from participants who describe the cookie theft picture from the Boston diagnostic aphasia examination [17] , the dataset encompasses speakers with different conditions for speaker heterogeneity and transcribed speech for ASR tasks, making it an excellent resource for exploring speaker heterogeneity issues in ASR. Utterances within each session were segmented, and those lasting less than 0.1 seconds were removed.

2.2. CharDiv-clustered FL strategy

2.2.1. *Federated learning with clusters*

The training process begins with K-means model training for data clustering, followed by individual client ASR models training using FL. The overall procedure is outlined in Algorithm 1.

During the K-means model (KM) training phase, the server first sends W_0^G to each client. W_0^G is obtained by fine-tuning a pre-trained ASR model on the server's data to enhance its ability to capture both elderly and AD speech characteristics, improving the quality of the clustering metric. The notation G signifies "global", and 0 denotes the initial model. Each client c uses W_0^G to extract CharDiv for each sample from its dataset D_c, and returns the resulting CharDiv embeddings E_c to the server. The server utilizes E_c from all clients to generate KM.

Subsequently, ASR models undergo a four-step training process for each client. Initially, all clients receive the same model weights W_0^G from the server as the initial model weights

[1] https://github.com/Victoria-Wei/Cluster-based-Personalized-Federated-Learning-with-CharDiv.

Algorithm 1 Federated learning with clusters

W stands for model weights; G stands for global model; C stands for the number of clients; D stands for training sets; E stands for CharDiv embeddings; KM stands for K-means model; K stands for the number of clusters, and R is the number of FL rounds
procedure FEDERATEDLEARNINGWITHCLUSTERS
  Initialize W_0^G (using the proxy data)
  Train Kmeans:
  for each client c in parallel **do**
    $E_c \leftarrow ExtractCharDiv(W_0^G, D_c)$
  end for
  $KM \leftarrow KM.fit((E_1, E_2, ..., E_C))$
  Train model:
  for each cluster k in parallel **do**
    $W_0^{G,k} \leftarrow W_0^G$
  end for
  for r=1,2,...,R **do**
    for each client c in parallel **do**
      for each cluster k in parallel **do**
        $D_{c,k} \leftarrow ClusterData(KM, D_c)$
        $W_r^{c,k} \leftarrow localTraining(W_{r-1}^{G,k}, D_{c,k})$
      end for
    end for
    $W_r^{G,k} \leftarrow \sum_{c=1}^{C} \frac{1}{C} W_r^{c,k}$
  end for
end procedure

for each of the K models, where K denotes the number of clusters. Then, each client clusters its data D_c into K segments using KM based on CharDiv derived from W_0^G, labeled as $D_{c,k}$ with k representing the corresponding cluster. This process is shown as the "data distributor" in Figure 1. Each client uses each of its data segments, to train a cluster-specific model, yielding weights $W_r^{c,k}$ for round r, which are returned to the server when local training is completed. At the end of each round, the server aggregates model weights from all clients by simple averaging over model weights trained with data of the same cluster label, resulting in K distinct sets of aggregated model weights $W_r^{G,k}$, which act as the initial model weights for the next round. The process can be terminated after a certain number of rounds, or when $W_r^{G,k}$ converge.

This process yields a coalition of K ASR models for each

client, and all clients share the same ASR coalition for inference. During inference, an input utterance is first distributed into its cluster by KM using CharDiv derived from W_0^G and is then decoded by the corresponding ASR model.

2.2.2. Character diversity (CharDiv)

CharDiv is designed to capture the text token distribution, including word usage and prosody differences, of each utterance without revealing the actual text content. For n-th utterance u_n, the ASR decoding process generates a raw character list, denoted as $u_n = \{x_1, x_2, ...x_T\}$, where $x_i \in \{a, b, c, ..., < unk >\}$ is the output character for time-step i. The set includes 26 English letters, along with 6 characters indicating pause, unknown, and others. T is the length of the utterance in time steps. For each possible character x_i, its frequency is calculated as N_{x_i}/T, where N_{x_i} is the occurrence of character x_i in the raw character list. Frequencies of 32 possible characters are then sorted to form $CharDiv_n = \{freq_{c_1}, freq_{c_2}, ...freq_{c_D}\}$, where c_i is the character with the i-th largest frequency $freq_{c_i}$, and D is the number of possible characters ($D = 32$). This vector signifies the diversity of characters spoken in the utterance, referred to as "character diversity" or CharDiv in short.

3. Results and analysis

3.1. Experimental settings

3.1.1. Data splits

To facilitate FL, the original dataset underwent restructuring, with data divided into distinct groups. A speaker-wise split was performed to ensure no overlapping speakers among each group, creating a setting with one server and five clients, each holding a portion of data. The server's data contains 50% of ADReSS training data, while the remaining 50% is distributed among the five clients. Within client data, an utterance-wise split allocated 70% for training, 10% for validation, and 20% for testing. Once the hyperparameters are decided, validation data is merged into the training data for all experiments. In addition to the random setting, a diverse scenario with highly heterogeneous data is explored, where the distribution of healthy controls (HC) and AD speakers differs among all clients. The demographics of the server and clients are presented in Table 1.

3.1.2. Comparison methods

We compared our method with various models in diverse and random client settings. The models, including 2 baseline methods, 3 non-FL methods, and 6 FL methods, are as follows:

- Baseline methods
  Pre-trained ASR [18]: data2vec ASR model with the pre-training setting of 'data2vec-audio-large-960h'.
  Fine-tuned ASR: fine-tuning the **Pre-trained ASR** model on server's data.
- Non-FL methods
  Centralized training: fine-tuning the **Fine-tuned ASR** model on all the client data at a central server.
  Fine-tuned clients: fine-tuning the **Fine-tuned ASR** model locally for each client.
  Fine-tuned speakers: fine-tuning the **Fine-tuned ASR** model locally for each speaker.
- FL methods
  FL: FedAvg [19] with simple averaging in aggregation.
  Weighted FL: FedAvg in [19].

Table 1: *Demographics of data among server and clients. (HC: healthy controls, AD: Alzheimer's diseased)*

	Client setting	Server	Clients				
			Client 1	Client 2	Client 3	Client 4	Client 5
People	Diverse	AD: 27 HC: 27	AD: 10 HC: 0	AD: 9 HC: 3	AD: 5 HC: 5	AD: 3 HC: 9	AD: 0 HC: 10
	Random		AD: 3 HC: 3	AD: 6 HC: 8	AD: 7 HC: 5	AD: 6 HC: 4	AD: 5 HC: 7
Utterance	Diverse	906	216	260	154	175	157
	Random		94	198	285	186	199

FedProx [10]: generalized FL regulating client weights by additional constraint loss during local training.
FML [11]: personalized FL with mutual model for information exchange and local model.
CBFL [13]: cluster-based FL modified to our case, using ASR embeddings as clustering metric.
CPFL-emb.: with ASR embeddings as clustering metric.

3.1.3. Training parameters

Our ASR system, based on the data2vec end-to-end framework [18], comprises 313,308,192 parameters and employs standard connectionist temporal classification (CTC) loss in training. Clients perform local training for 10 epochs, and the number of FL rounds (R) is set to 10. The training process involves $K = 7$ clusters and $C = 5$ clients for both diverse and random settings. A grid search for hyperparameter K on the validation set selects $K = 7$ based on the lowest WER among candidate numbers. Experiments are conducted on DGX Station A100 using a single GPU, with CPFL training taking up to $38\,\mathrm{h}$, varying with different settings. During multiple tests with audio files of various lengths, we observed an average inference time of $2\,\mathrm{s}$, with an average audio file duration of $3.7\,\mathrm{s}$.

3.2. Superiority of CPFL and CharDiv

3.2.1. Superiority of FL methods over non-FL ones

In assessing FL's advantages over traditional fine-tuning for biased and limited AD speech, we compare non-FL and FL methods. All three non-FL settings show performance drops compared to **Fine-tuned ASR** in Table 2. **Centralized training** has a WER of 46.6%. **Fine-tuned clients** experiences WER increases to 48.58% and 50.11% for diverse and random client settings, respectively. These declines may stem from heterogeneity within each client's training data. Even **Fine-tuned speakers** worsens the WER to 39.71% due to limited speaker data and high intra-speaker heterogeneity. Surprisingly, the least effective FL method outperforms non-FL methods, showcasing FL's significant potential, as highlighted in [6], for training models with limited yet biased data. A robust FL method is important for handling highly biased and limited data.

3.2.2. Superiority of cluster-based FL over model-based FL

To show the potential of cluster-based FL, we reveal the limitations of model-based FL methods by comparing them to vanilla FL approaches. First, **FedProx**, a generalized model-based FL method addressing client heterogeneity, performs similarly to vanilla FL methods (**FL** and **Weighted FL**) with WERs of about 34%. This implies that adjusting high-level model weights may not suffice for handling the high heterogeneity in our dataset and that a universally applicable model may not fit all samples. Second, in FML, a personalized model-based

Table 2: *ASR performances in WER (%) of different models*

		Diverse	Random
Baseline	**Pre-trained ASR**	56.08	
	Fine-tuned ASR	35.12	
Non-FL	**Centralized training**	46.60	
	Fine-tuned speakers	39.71	
	Fine-tuned clients	48.58	50.11
Vanilla FL	**FL**	33.59	34.89
	Weighted FL	34.58	34.12
Generalized FL	**FedProx**	34.05	34.35
Personalized FL	**FML (local)**	38.33	37.41
	FML (mutual)	33.44	32.75
Cluster-based FL	**CBFL**	34.51	34.66
	CPFL – emb	33.21	32.59
	CPFL – CharDiv	**29.99**	**30.22**

FL approach, local models (**FML (local)**) consistently underperform vanilla FL models, with WER increases of over 2.5% across different client settings, which indicates the challenges in achieving optimal performance with personalized model-based FL when facing high within-client heterogeneity. Cluster-based methods, which reduce within-cluster heterogeneity by grouping similar samples or clients, offer a promising solution for tackling the complexities of highly heterogeneous datasets.

3.2.3. Superiority of CharDiv-clustered CPFL

In order to demonstrate the suitability of our CPFL and CharDiv design in FL AD-oriented ASR tasks, we compare it with CBFL and explore the impact of different clustering metrics. Despite the promise of cluster-based FL, CBFL's approach of training models on all samples for different models may harm model performance. Instead, we use corresponding data for separate models. With the same clustering metric, **CPFL-emb** shows WER improvements of 1.3% and 2.07% for diverse and random client settings, highlighting its superiority over CBFL. The designed clustering metric, CharDiv, is more effective in capturing characteristics in speech data, especially for text token information, forming more robust clusters. **CPFL-CharDiv** achieves the lowest WER of 29.99% and 30.22% for diverse and random client settings, respectively. The WER improvements are 3.6% and 4.67% compared to **FL**, and 3.22% and 2.37% compared to **CPFL-emb**. The results show that the design for CPFL and CharDiv contributes to enhanced model performance.

3.3. Cluster analysis

To evaluate our method's impact on resolving heterogeneity, we analyze each cluster's characteristics. Table 3 compares WER per cluster between CPFL-CharDiv models and the vanilla FL model, revealing notable WER improvements by our method, especially in clusters 2 and 6. Further analysis of the CharDiv distribution reveals distinct patterns in these clusters compared to others. Table 4 highlights the first dimension from CharDiv's 32 dimensions to illustrate the primary differences. Notably, clusters 2 and 6 both exhibit elevated values in this dimension, corresponding to "<pad>," a special token denoting pauses predicted by the data2vec model, across all clusters. We further investigate pause distributions, categorizing samples into "long-pause" (over 80% predicted time steps as "<pad>"), "short-pause" (less than 60%), and "medium-pause" (in between). Table 4 depicts the distribution of these pause types in each cluster, revealing that clusters 2 and 6 exclusively consist of "long-pause" samples. Conversely, cluster 1 and 5 solely comprise "short-pause" samples and only cluster 7 has less than 20% of its data as "long-pause," showing variations in pause usage among clusters. The following samples demonstrate raw output tokens, with "$(*N)$" indicating N repetitions of the previous token. A sample from cluster 5, "<pad>(*7)IID<pad>ON"TKNOW<pad>(*2)," exhibits a small number of "<pad>", while another sample from cluster 2, "<pad>(*32) O<pad>(*5) K<pad>(*2) A<pad>(*5) Y<pad>(*9)," exhibits a large number. Both samples are from the same speaker, highlighting varying pause usage due to different speaking conditions, emphasizing the importance of using sample as clustering unit for improved clusters. The findings underscore the noticeable impact of a speaker's pause usage on ASR model learning and the effectiveness of our method in addressing scenarios where speakers use longer pauses while uttering fewer words. Heterogeneity in pause usages seems to be a key heterogeneity that is evident in this cohort.

Table 3: *Comparisons of WER (%) in FL and CPFL-CharDiv*

	FL	CPFL-CharDiv	WER reduction
Cluster 1	17.02	18.09	-1.07
Cluster 2	**54.17**	**45.00**	**9.17**
Cluster 3	33.06	28.93	4.13
Cluster 4	28.84	28.09	0.75
Cluster 5	16.42	14.93	1.49
Cluster 6	**44.54**	**38.86**	**5.68**
Cluster 7	34.84	30.32	4.52

Table 4: *Cluster distributions of sentences different in pause length and CharDiv's first dimension*

	Sample distribution (%)			Average value of 1st-dim CharDiv (%)
	Long	Medium	Short	
Cluster 1	0	0	100	51.13 ± 3.00
Cluster 2	**100**	**0**	**0**	**95.08 ± 2.51**
Cluster 3	0	59.46	40.54	60.76 ± 2.55
Cluster 4	0	100	0	69.11 ± 2.38
Cluster 5	0	0	100	37.69 ± 4.93
Cluster 6	**100**	**0**	**0**	**86.36 ± 2.53**
Cluster 7	15.49	84.51	0	77.32 ± 2.45

4. Conclusions

This study introduces a cluster-based FL method for AD-oriented ASR, featuring the designed clustering metric CharDiv and an ASR system based on clusters to mitigate text token heterogeneity. We show FL's superiority on limited yet biased data, enhance cluster-based FL through the CPFL strategy, and use CharDiv to capture ASR output token distributions. Analyzing per-cluster WER improvements and CharDiv distributions reveals reduced pause usage heterogeneity, benefiting ASR training. While our focus here is on a subset of text token heterogeneity, further exploration is needed for other AD-related variations, such as part-of-speech usage or vocalization differences.

5. References

[1] T. Wang, J. Deng, M. Geng, Z. Ye, S. Hu, Y. Wang, M. Cui, Z. Jin, X. Liu, and H. Meng, "Conformer Based Elderly Speech Recognition System for Alzheimer's Disease Detection," in *Proc. Interspeech 2022*, 2022, pp. 4825–4829.

[2] Y. Wang, J. Deng, T. Wang, B. Zheng, S. Hu, X. Liu, and H. Meng, "Exploiting prompt learning with pre-trained language models for alzheimer's disease detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] M. Zvěřová, "Clinical aspects of alzheimer's disease," *Clinical biochemistry*, vol. 72, pp. 3–6, 2019.

[4] K. López-de Ipiña, J.-B. Alonso, C. M. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage *et al.*, "On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis," *Sensors*, vol. 13, no. 5, pp. 6730–6745, 2013.

[5] Z. Ye, S. Hu, J. Li, X. Xie, M. Geng, J. Yu, J. Xu, B. Xue, S. Liu, X. Liu *et al.*, "Development of the cuhk elderly speech recognition system for neurocognitive disorder detection using the dementiabank corpus," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6433–6437.

[6] Z. Xiong, Z. Cheng, X. Lin, C. Xu, X. Liu, D. Wang, X. Luo, Y. Zhang, H. Jiang, N. Qiao *et al.*, "Facing small and biased data dilemma in drug discovery with enhanced federated learning approaches," *Science China Life Sciences*, pp. 1–11, 2021.

[7] Y. Gao, T. Parcollet, S. Zaiem, J. Fernandez-Marques, P. P. de Gusmao, D. J. Beutel, and N. D. Lane, "End-to-end speech recognition from federated acoustic models," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7227–7231.

[8] S. S. Azam, T. Likhomanenko, M. Pelikan *et al.*, "Importance of smoothness induced by optimizers in fl4asr: Towards understanding federated learning for end-to-end asr," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[9] K. Nandury, A. Mohan, and F. Weber, "Cross-silo federated training in the cloud with diversity scaling and semi-supervised learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3085–3089.

[10] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[11] T. Shen, J. Zhang, X. Jia, F. Zhang, Z. Lv, K. Kuang, C. Wu, and F. Wu, "Federated mutual learning: a collaborative machine learning method for heterogeneous data, models, and objectives," *Frontiers of Information Technology & Electronic Engineering*, vol. 24, no. 10, pp. 1390–1402, 2023.

[12] B. Farahani, S. Tabibian, and H. Ebrahimi, "Towards a personalized clustered federated learning: A speech recognition case study," *IEEE Internet of Things Journal*, 2023.

[13] L. Huang, A. L. Shea, H. Qian, A. Masurkar, H. Deng, and D. Liu, "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records," *Journal of biomedical informatics*, vol. 99, p. 103291, 2019.

[14] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge," in *Proc. Interspeech 2020*, 2020, pp. 2172–2176.

[15] A. C. Morris, V. Maier, and P. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition," in *Eighth International Conference on Spoken Language Processing*, 2004.

[16] A. M. Lanzi, A. K. Saylor, D. Fromm, H. Liu, B. MacWhinney, and M. L. Cohen, "Dementiabank: Theoretical rationale, protocol, and illustrative analyses," *American Journal of Speech-Language Pathology*, vol. 32, no. 2, pp. 426–438, 2023.

[17] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston diagnostic aphasia examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[18] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[19] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.