Personalized Federated Learning with Fuzzy Clustering for Dysarthric Speech Recognition

Jie-Shiang Yang
dept. Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan
snoopy963852@gapp.nthu.edu.tw

Jing-Tong Tzeng
dept. Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan
roger37890426@gapp.nthu.edu.tw

Chi-Chun Lee
dept. Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan
cclee@ee.nthu.edu.tw

***Abstract*—Pathological speech recognition is challenging because clinical datasets are scarce, variable, and subject to strict privacy constraints preventing cross-institutional data sharing. These regulations necessitate federated learning (FL) for collaborative training without sharing raw data. However, FL degrades under non-IID data. Hard-clustering FL addresses this by partitioning clients into groups but imposes rigid boundaries, discards boundary samples, and suffers performance drops as cluster numbers increase. We propose Fuzzy Cluster-Based Personalized Federated Learning (FCPFL), using fuzzy C-means to softly group clients and pseudo-label-guided feature selection to identify discriminative features. FCPFL weights client updates by membership degree, allowing boundary samples to participate in multiple clusters and increasing training data by 25%. Experiments show FCPFL reduces word error rate (WER) by 4.82% and 1.74% on ADReSS and TORGO, compared to hard-clustered FL baselines.**

***Index Terms*—Federated learning, pathological speech recognition, fuzzy clustering, feature selection, Alzheimer's disease, dysarthria, automatic speech recognition**

I. Introduction

Automatic speech recognition (ASR) excels on standard benchmarks [1]–[3], but on pathological speech, word error rates are 2–5× higher due to articulatory imprecision, prosodic irregularities, and atypical acoustics [4]–[6]. Deploying ASR in clinical settings faces two challenges. First, privacy: clinical speech is spread across hospitals with regulations preventing raw-data sharing [7]–[9]. Second, heterogeneity: pathological speech varies by disorder (e.g., Alzheimer's, dysarthria), severity, and recording conditions, yielding highly non-IID data distributions that hinder generalization [8], [10].

Federated Learning (FL) is particularly well-suited for clinical speech applications because strict privacy regulations prohibit sharing raw audio or transcripts across hospitals and research institutions. By exchanging only model updates, FL enables collaborative training without exposing sensitive data [7], [8], [11], [12]. However, standard FL algorithms such as FedAvg [1] and FedProx [13] still struggle when client data are non-IID, exhibiting slower convergence and higher error rates as label distributions diverge [1], [13]–[15].

Because pathological speech data vary not only by disorder type but also by disease progression and recording conditions, several strategies have emerged to mitigate heterogeneity. FedProx adds a proximal term to mitigate client drift but cannot capture multimodal cluster structures in disease-affected speech [13], [14], [16]. Personalized FL methods like FedMeta improve local adaptation but often ignore continuous severity gradients and offer limited clinical insights [10], [17], [18].

As an alternative, hard clustering–based FL methods group clients with similar characteristics for separate aggregation. In medical imaging, several studies have partitioned hospitals [19] or imaging devices [20] into fixed clusters to mitigate domain shifts. These approaches deliver benefits when the chosen cluster count matches true data heterogeneity but perform poorly if it does not [7]. Similarly, in disordered speech recognition, Hsu *et al.* applied the CharDiv metric to cluster dysarthric clients and achieved modest WER improvements on ADReSS [4], [21]. Nevertheless, hard clustering presupposes a fixed number of clusters, enforces rigid boundaries that discard samples near cluster edges, and relies on an English-specific metric—restricting its applicability across languages and diverse pathological conditions [4], [5], [10]. Therefore, a soft clustering mechanism that preserves boundary information and privacy remains necessary.

To address these challenges, we propose Fuzzy Cluster–Based Personalized Federated Learning (FCPFL). FCPFL uses Fuzzy C-Means (FCM) to compute soft membership degrees over client feature distributions, allowing each client's update to be weighted by its similarity to multiple cluster centroids and thus providing smooth adaptation to severity gradients [22]–[24]. Concurrently, we perform FCM-weighted pseudo-label feature selection. Moreover, we introduce membership scores guide pseudo-labels assignment to unlabeled feature vectors, enabling identification of a compact set of highly discriminative acoustic–linguistic features for each soft cluster. This leverages the success of pseudo-labeling in semi-supervised audio tasks [25]–[27].

Our work makes three main contributions. First, we introduce a soft-clustering personalization framework in which a fuzzy-membership aggregation scheme captures continuous heterogeneity without requiring a fixed number of clusters. Second, we develop an FCM-guided pseudo-label feature-selection procedure that dynamically extracts feature subsets tailored to each soft cluster, thereby enhancing both interpretability and model efficiency [28]. Third, we conduct comprehensive federated experiments on the ADReSS [21] and TORGO [29] corpora, demonstrating relative WER reductions of 4.82 % and 1.74 % compared to hard-clustered FL baselines.

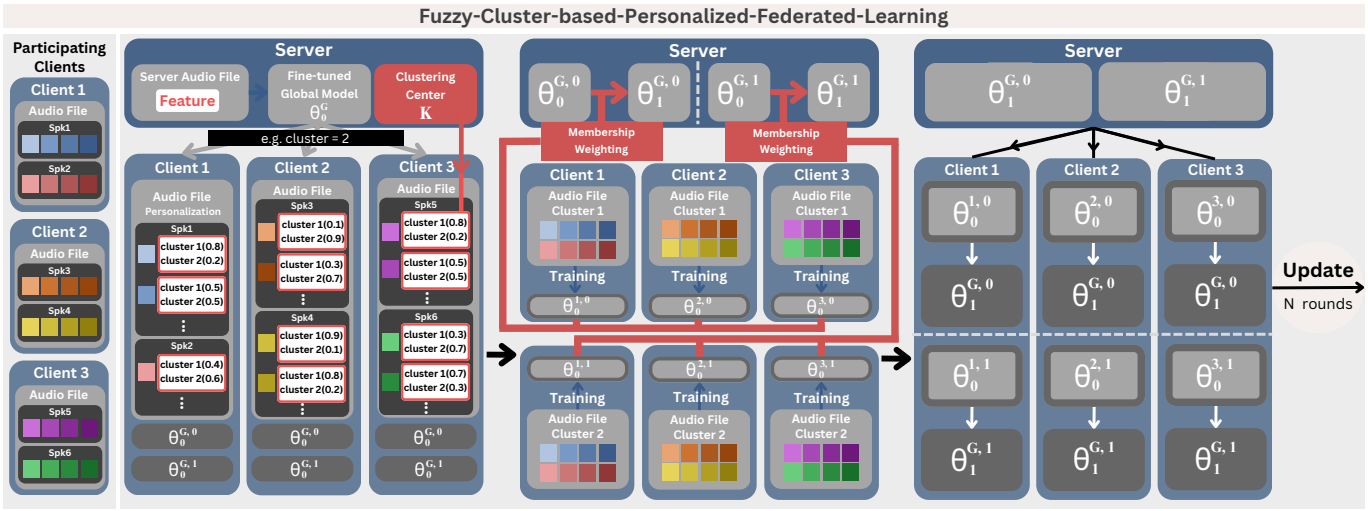

Fig. 1. The Fuzzy Cluster-Based Personalized Federated Learning (FCPFL) framework architecture.

II. Related Works

A. Federated Learning for ASR

Federated Learning (FL) enables collaborative ASR training without sharing raw data [30], [31]. However, under severe non-IID conditions—such as differing pathology severities—standard algorithms like FedAvg suffer from degraded convergence and reduced accuracy [8], [14], [32].

B. Hard Clustering FL for Pathological Speech

To address the extreme heterogeneity of pathological speech, several federated learning approaches partition clients into disjoint clusters and train separate models for each group. While Clustered Personalized Federated Learning (CPFL) [4] clusters dysarthric speakers using CharDiv and demonstrates modest WER reductions on ADReSS, its effectiveness is constrained by a reliance on predefined cluster counts, rigid group boundaries, and language-specific metrics.

Similarly, Li *et al.* [33] proposed a hierarchical multi-cluster FL framework (FedMP) to handle diverse client populations. Although FedMP improves personalization by assigning each client to a single hard cluster, it does not account for severity levels in pathological data, which limits flexibility and results in unstable performance as the number of clusters increases. These issues highlight the fragility of hard clustering under non-IID conditions, particularly when client characteristics span continuous spectrums or evolve over time [34].

C. Fuzzy Clustering in FL and Medical Domains

Fuzzy C-Means (FCM) overcomes hard boundaries by assigning soft membership degrees, allowing samples to belong to multiple clusters. Hu *et al.* applied FCM within FL to handle longitudinal non-IID behavioral data; by allowing samples to belong to multiple clusters and capturing boundary samples that hard clustering would overlook, they achieved higher classification accuracy than traditional hard clustering [35]. Wang *et al.*'s FedRFC uses recursive fuzzy clustering so clients can participate in multiple clusters, outperforming hard clustering on synthetic benchmarks [36]. In medical imaging, Zhang *et al.*'s SplitAVG applies fuzzy-like weighting to multi-hospital data, yielding stable convergence and 3% segmentation accuracy gain [37]. However, no work has combined FCM with speech-specific feature selection for pathological ASR.

D. Motivation and Gap

Although FL has been applied to mildly non-IID healthy speech [38] and hard clustering methods have been explored for pathological ASR [4], [19], these assume discrete severity partitions that fail to capture continuous transitions. FCM has shown promise in other FL and medical imaging tasks [35], [36], yet not in speech. Pseudo-labeling aids semi-supervised ASR [25], [26], but has not been integrated with fuzzy clustering to jointly address privacy and extreme heterogeneity in pathological ASR. FCPFL addresses this gap by leveraging FCM to preserve boundary information and pseudo-label–guided feature selection for privacy and heterogeneity handling, validated on ADReSS and TORGO.

III. Methodology

A. Fuzzy Cluster-Based Personalized Federated Learning

Figure 1 illustrates our FCPFL framework, which employs soft clustering to model the continuous nature of pathological speech variations. Unlike hard clustering approaches, FCPFL computes membership degrees that reflect gradual transitions between different pathological characteristics.

*1) **Global Feature Collection and Cluster Center Discovery**:* The server collects representative feature matrices from participating clients and applies Fuzzy C-Means (FCM) clustering [22] to discover c cluster centers. FCM partitions the feature space into overlapping clusters by minimizing:

$$J_m(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^m \|\mathbf{y}_i - \mathbf{v}_j\|_A^2 \quad (1)$$

where $\mathbf{U} = [u_{ij}]$ is the $c \times N$ fuzzy partition matrix with u_{ij} representing the membership degree of data point i to cluster j, $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_c\}$ are the cluster centers, m is the fuzziness parameter ($1 < m < \infty$), and $\|\cdot\|_A$ denotes the A-norm with positive definite matrix A. A represents the identity matrix for standard Euclidean distance in our implementation.

The membership degrees are updated using:

$$u_{ij} = \frac{1}{\sum_{l=1}^{c} \left(\frac{\|\mathbf{y}_i - \mathbf{v}_j\|_A}{\|\mathbf{y}_i - \mathbf{v}_l\|_A} \right)^{\frac{2}{m-1}}} \tag{2}$$

The cluster centers are updated using:

$$\mathbf{v}_j = \frac{\sum_{i=1}^{N} u_{ij}^m \mathbf{y}_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{3}$$

The optimal cluster count c^* is determined by minimizing the Xie–Beni index [39]:

$$c^* = \arg \min_{c \in \{2,\ldots,c_{\max}\}} \frac{\sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^m \|\mathbf{y}_i - \mathbf{v}_j\|^2}{N \cdot \min_{j \neq l} \|\mathbf{v}_j - \mathbf{v}_l\|^2} \tag{4}$$

*2) **Membership-Weighted Local Training**:* Each client k maintains c local models $\{\boldsymbol{\theta}_{k,1}, \boldsymbol{\theta}_{k,2}, ..., \boldsymbol{\theta}_{k,c}\}$ corresponding to discovered clusters. Clients perform membership-weighted training using:

$$\mathcal{L}_{\text{weighted}} = \sum_{i=1}^{n_k} \sum_{j=1}^{c} u_{ij} \cdot \ell(\mathbf{x}_i, y_i; \boldsymbol{\theta}_{k,j}) \tag{5}$$

where $\ell(\mathbf{x}_i, y_i; \boldsymbol{\theta}_{k,j})$ is the Connectionist Temporal Classification (CTC) loss function for cluster j with model parameters $\boldsymbol{\theta}_{k,j}$ [40]. To ensure training efficiency, only samples with membership degrees above threshold M participate in training: $\mathcal{D}_{k,j} = \{\mathbf{x}_i \in \mathcal{D}_k \mid u_{ij} \geq M\}$.

*3) **Membership-Weighted Server Aggregation**:* Following the FedAvg aggregation principle [1], the server aggregates client models using membership-weighted averaging:

$$\boldsymbol{\theta}_{\text{global}}^{(j)} = \frac{\sum_{k=1}^{K} w_{k,j} \cdot \boldsymbol{\theta}_k^{(j)}}{\sum_{k=1}^{K} w_{k,j}} \tag{6}$$

where $w_{k,j} = \sum_{i \in \mathcal{D}_{k,j}} u_{ij}$ represents the total membership weight contributed by client k to cluster j.

*4) **Multi-Model Fusion for Inference**:* During inference, the system employs a membership-weighted ensemble among valid clusters:

$$\hat{y} = \arg\max_{y} \sum_{j: u_j \geq M} \frac{u_j}{\sum_{l: u_l \geq M} u_l} \cdot P_j(y|\mathbf{x}) \tag{7}$$

where $P_j(y|\mathbf{x})$ is the prediction probability of cluster j's ASR model for input $\mathbf{x}$.

B. FCM-Guided Pseudo-Label Feature Selection

The discriminative quality of cluster partitions fundamentally determines the effectiveness of FCPFL. Suboptimal clustering leads to inappropriate model specialization, ineffective parameter aggregation, and degraded inference performance. To address these challenges, we develop a specialized feature selection methodology that ensures FCM receives the most discriminative feature representation.

*1) **Feature Extraction and Categorization**:* We extract 11 validated features for speech-disorder tasks, grouped into 5 categories that reflect different aspects of pathological speech. Acoustic features include fundamental frequency standard deviation, mel-frequency cepstral coefficients mean, energy mean, and energy standard deviation, which reflect physiological changes in speech production. Pause features comprise pause rate, which is calculated as pause duration divided by total duration, pause mean as the average pause duration, and pause count, indicating cognitive processing deficits. Linguistic features cover speech-rate variance and word rate in words per unit time, capturing language production difficulties. We also include pronunciation features via voiced rate, defined as the proportion of voiced segments, and deep representation features extracted from Data2vec-audio-large hidden-layer activations [41], which provide learned acoustic-linguistic representations.

*2) **Two-Stage Selection Framework**:* Our feature selection operates as a Filter+Embedded framework enhanced with FCM membership weighting to address the absence of healthy controls in pathological-only datasets.

Filter Stage: The filter stage evaluates features independently of any specific learning algorithm by analyzing their statistical properties and relevance to cluster separation. This stage efficiently eliminates irrelevant features before computationally expensive model training. We compute Weighted Mutual Information (WMI) for each feature j using Eq. (8), then select the top K_f features with the highest WMI scores:

$$\text{WMI}_j = \sum_{i=1}^{N} \sum_{l=1}^{c^*} u_{il} \cdot I(X_{i,j}; l) \tag{8}$$

where $I(X_{i,j}; l)$ represents the mutual information between feature j of sample i and cluster l, computed as $I(X_{i,j}; l) = \sum_{x,c} p(x,c) \log \frac{p(x,c)}{p(x)p(c)}$ where $p(x,c)$ is the joint probability distribution of feature values and cluster assignments. This weighting mechanism evaluates features based on their contribution to fuzzy cluster separation.

Embedded Stage: The embedded stage performs feature selection within the context of a specific learning algorithm, considering feature interactions and their collective predictive power. Unlike filter methods, embedded approaches can capture feature dependencies and non-linear relationships. We train weighted L1-regularized logistic regression using Eq. (9), where the L1 penalty induces sparsity for feature selection:

$$\min_{\mathbf{W}} \left[-\sum_{i=1}^{N} \sum_{l=1}^{c^*} u_{il} \log \frac{e^{\mathbf{w}_l^\top \mathbf{x}_i}}{\sum_{j=1}^{c^*} e^{\mathbf{w}_j^\top \mathbf{x}_i}} + \lambda \sum_{j,l} |W_{jl}| \right] \tag{9}$$

where $\mathbf{W}$ is the weight matrix, $\mathbf{w}_l$ is the weight vector for cluster l, λ is the L1 regularization parameter, and W_{jl} represents the weight of feature j for cluster l. The L1 regularization drives unimportant feature weights toward zero for automatic selection.

After M repetitions, we select features with frequency $\geq \theta$ and rank them by average importance scores.

C. Complete Algorithm Specifications

The FCPFL algorithm consists of three main stages. First, global pre-training obtains a base ASR model θ_0^G. Second, FCM(c^*, m) clustering on extracted features yields cluster centers and membership matrix $\mathbf{U}$. Each cluster's global model initializes from θ_0^G. Third, federated learning: selected clients filter data by threshold M, perform membership-weighted training (Eq. (5)), and send updates. Server aggregates using weighted averaging (Eq. (6)).

Algorithm 1 FCPFL Federated Training

Require: Datasets $\{\mathcal{D}_k\}_{k=1}^K$, model θ_0^G, threshold M, clusters c^*, rounds T, epochs E, fraction frac
Ensure: $\{\boldsymbol{\theta}_T^{G,1}, \ldots, \boldsymbol{\theta}_T^{G,c^*}\}$

Initialize $\boldsymbol{\theta}_0^{G,j} \leftarrow \theta_0^G$ for $j = 1, \ldots, c^*$
for $t = 1$ to T **do**
  $S_t \leftarrow \text{RandomSample}(K, \max(\lfloor \text{frac} \times K \rfloor, 1))$
  for $j = 1$ to c^* **do**
    Initialize: $\mathcal{W}^j \leftarrow \emptyset$, $\mathcal{M}^j \leftarrow \emptyset$
    for each client $k \in S_t$ **do**
      $\mathcal{D}_{k,j} \leftarrow \{\mathbf{x}_i \in \mathcal{D}_k \mid u_{ij} \geq M\}$
      if $\mathcal{D}_{k,j} \neq \emptyset$ **then**
        $w_{k,j} \leftarrow \sum_{i \in \mathcal{D}_{k,j}} u_{ij}$
        $\boldsymbol{\theta}_k^{(j)} \leftarrow \text{LocalTrain}(\mathcal{D}_{k,j}, \boldsymbol{\theta}_{t-1}^{G,j}, E)$
        $\mathcal{W}^j \leftarrow \mathcal{W}^j \cup \{\boldsymbol{\theta}_k^{(j)}\}$, $\mathcal{M}^j \leftarrow \mathcal{M}^j \cup \{w_{k,j}\}$
      end if
    end for
    if $\mathcal{W}^j \neq \emptyset$ **then**
      $\boldsymbol{\theta}_t^{G,j} \leftarrow \frac{\sum_k w_{k,j} \cdot \boldsymbol{\theta}_k^{(j)}}{\sum_k w_{k,j}}$
    else
      $\boldsymbol{\theta}_t^{G,j} \leftarrow \boldsymbol{\theta}_{t-1}^{G,j}$
    end if
  end for
end for
return $\{\boldsymbol{\theta}_T^{G,1}, \ldots, \boldsymbol{\theta}_T^{G,c^*}\}$

IV. Experimental Settings

A. Datasets

This research employs two publicly available pathological speech corpora: ADReSS and TORGO. Both datasets require access authorization, highlighting the privacy constraints motivating our federated learning approach.

The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge dataset [21] consists of 156 picture-description recordings (78 Alzheimer's patients, 78 healthy controls) from the DementiaBank repository [42], balanced for age and gender. The TORGO database [29] comprises aligned acoustic and articulatory recordings from eight dysarthric speakers (5 male, 3 female; cerebral palsy or ALS) and matched controls, collected by the University of Toronto and Holland-Bloorview Kids Rehab Hospital [43].

TABLE I
Overall WER Performance Comparison

Method Category	ADReSS	TORGO
Baseline Methods		
Pre-trained ASR	0.6869	0.7887
Fine-tuned ASR	0.4177	0.5016
Standard Federated Learning		
FedAvg	0.4054	0.2633
Weighted FL	0.4028	0.2386
Generalized FL		
FedProx	0.4054	0.2538
Parameter Efficient FL		
FedLoRA	0.3866	0.2614
FL-TAC	0.3901	0.2507
Hard Clustering FL		
CBFL	0.3887	0.2357
CPFL-embs	0.3925	0.2331
CPFL-CharDiv	0.3964	0.2358
Soft Clustering FL		
FCPFL	**0.3408**	**0.2157**

B. Experimental Setup

Dataset Partitioning: We exclusively use pathological speech data, excluding healthy controls, to focus on clinically relevant severity distinctions and avoid statistical imbalance from dominant healthy clusters. The datasets are split 50%/50% between server and clients, with the server portion used for global model initialization and cluster center determination.

Training Configuration: We employ Data2vec-audio-large-960h [41] as our backbone model, a general framework for self-supervised learning that predicts contextualized latent representations. Training parameters include a learning rate of 1e-5, 10 local epochs per federated round, and 10 total global epochs. Experiments are conducted on DGX Station A100[1].

Baseline Comparisons: We compare against non-federated methods (Pre-trained and Fine-tuned ASR), standard FL methods (FedAvg [1], Weighted FedAvg, FedProx [44]), parameter-efficient FL approaches (FedLoRA, FL-TAC [45]), hard clustering FL methods (CBFL, CPFL-embs, CPFL-CharDiv [4]), and our proposed soft clustering FCPFL.

Computational Overhead Analysis: Our approach indeed incurs higher communication overhead compared to standard federated learning due to the heterogeneous nature of pathological speech requiring specialized cluster-based models. However, this increased cost is justified by substantial performance improvements in clinical applications where accuracy is paramount.

[1]The implementation code is available at: https://github.com/Jie-shiang/Fuzzy-Cluster-based-Personalized-Federated-Learning.git

V. Results and Analysis

A. Overall Performance Analysis

Table I presents comprehensive WER comparisons across different method categories. Our proposed FCPFL achieves the best performance on both datasets, obtaining 0.3408 WER on ADReSS with a 4.82% reduction compared to the best hard clustering method CBFL at 0.3887, and 0.2157 WER on TORGO with a 1.74% reduction compared to CPFL-embs at 0.2331.These WER reduction translates directly to enhanced clinical utility for downstream diagnostic applications.

Baseline Performance: Pre-trained ASR models show high WER (ADReSS: 0.6869, TORGO: 0.7887), highlighting pathological speech recognition challenges. Fine-tuned ASR provides significant improvement but remains suboptimal.

Standard and Generalized FL: Traditional federated learning methods demonstrate substantial improvements over centralized fine-tuning, validating the effectiveness of collaborative learning for pathological speech.

Clustering-Based FL Advantages: Hard clustering methods (CBFL, CPFL variants) consistently outperform standard FL approaches, demonstrating the value of personalization strategies.

B. Feature Selection Analysis

Our FCM-guided feature selection methodology effectively identifies discriminative features through the Filter+L1 approach. For single features, mel-frequency cepstral coefficients mean (MFCC) consistently achieved the best performance, FCPFL: 0.3634/0.2243 vs. CPFL: 0.3814/0.2395 on ADReSS/TORGO. In multi-feature analysis shows the optimal TOP 3 combination significantly outperforms single-feature approaches, with FCPFL achieving 0.3408/0.2157 compared to CPFL's 0.3854/0.2367—relative improvements of 6.2% and 3.8% respectively.

Disease-Specific Feature Combinations: Through our Filter+L1 selection method, we identified optimal three-feature combinations reflecting pathological speech characteristics:

- **ADReSS:** MFCC + Pause_rate + Word_rate (reflecting cognitive decline impacts on language fluency)
- **TORGO:** MFCC + Voiced_rate + Energy_std (capturing motor control deficits in speech production)

These combinations demonstrate disease-specific feature importance: Alzheimer's speech prioritizes pause patterns and word production efficiency, while dysarthric speech emphasizes voicing control and energy variability, validating our methodology's ability to capture disorder-specific acoustic-linguistic characteristics. While our evaluation focuses on English datasets, our core features (f0 std, energy std, MFCC, pause patterns) are fundamentally physiologically-based and language-independent. Vocal fold dysfunction, respiratory patterns, and articulatory precision manifest consistently across languages, making our FCM clustering mechanism applicable to universal pathological speech characteristics regardless of linguistic content.

TABLE II
Statistical Analysis of FCPFL vs. CPFL

Dataset	Method	CV ↓	Correlation (r)	P-value
ADReSS	CPFL	0.0246	0.442	0.273
	FCPFL	0.0095	0.911	0.002
TORGO	CPFL	0.0165	0.345	0.403
	FCPFL	0.0116	0.810	0.015

C. Impact of Hyperparameter Variations on FCPFL

Both cluster granularity and membership threshold parameters fundamentally control the number of valid samples participating in training, making their analysis crucial for understanding FCPFL's robustness. To better comprehend FCPFL's advantage across different settings, we categorize pathological speech samples into three types:

- **Stable samples** (training frequency = 1): clear pathological characteristics consistently assigned to a single cluster.
- **Boundary samples** (training frequency > 1): fuzzy characteristics that contribute to multiple clusters.
- **Unutilized samples** (training frequency = 0): excluded due to low membership degrees.

As shown in Fig. 2, in hard clustering methods, WER begins to increase after $c = 4$ and degrades sharply by $c = 9$. Too many clusters can leave some clients with no valid samples, leading to overfitting and reduced generalization. In contrast, under soft clustering, increasing c has minimal effect on WER since virtually all client data participate in training and the weighted-loss mechanism in FCPFL mitigates performance fluctuations. The lower panel of Fig. 2 shows that FCPFL retains samples more steadily. While CPFL drops from 532 to 106 samples on ADReSS (an 80% reduction), FCPFL declines from 386 to 86 samples (78% reduction). Moreover, for ADReSS, the proportion of boundary samples increases from 23% to 38% as c grows from 2 to 9; across ADReSS and TORGO, the average gain in valid samples over clusters 2–9 is about 25%.

Fig. 3 reveals that threshold effects mirror cluster effects, and vice versa. The *Primary/Valid Ratio* is the ratio between original utterances (Primary) and total valid utterances (Valid)—if an utterance appears in multiple clusters, it is counted multiple times in the Valid count. Higher thresholds exclude more samples, while lower thresholds include potentially noisy data. The optimal range of 0.750–0.775 achieves the best balance, corresponding to 23–38% data augmentation via boundary samples. For both ADReSS and TORGO, the optimal threshold is around 0.75: too low invites noise and increases WER; too high excludes boundary samples, reducing training data. Both datasets exhibit similar optimal ranges, validating FCPFL's boundary-sample mechanism across Alzheimer's and dysarthria conditions.

Statistical validation through paired t-tests confirms significant improvements: ADReSS shows large effect size over CBFL (t = 4.12, $p < 0.001$, Cohen's d = 0.87), while TORGO

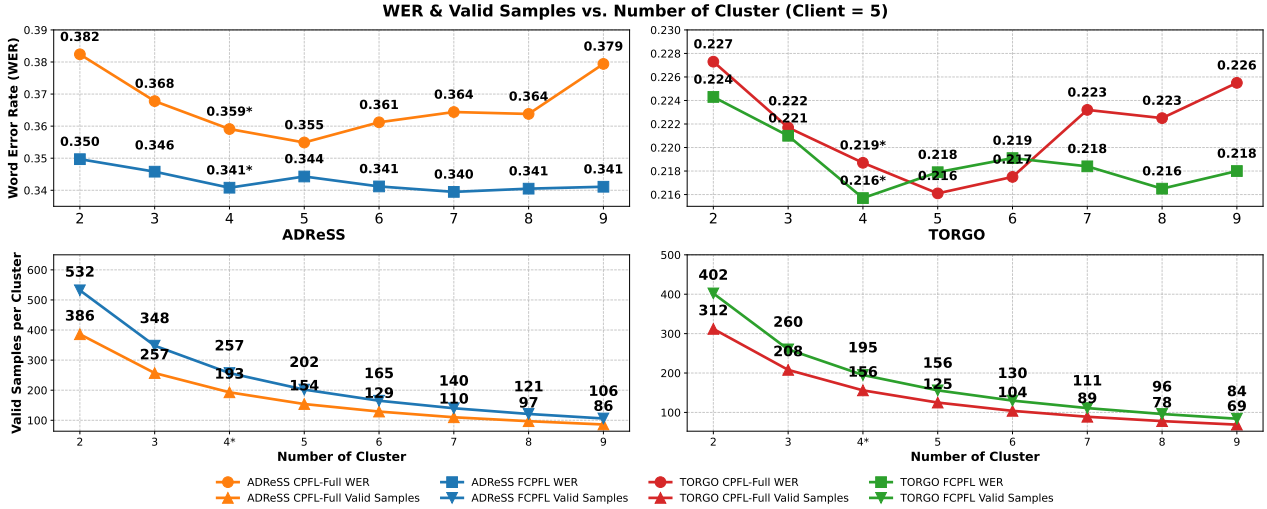

Fig. 2. Impact of cluster number on WER performance comparison between hard clustering (CPFL) and soft clustering (FCPFL) methods.

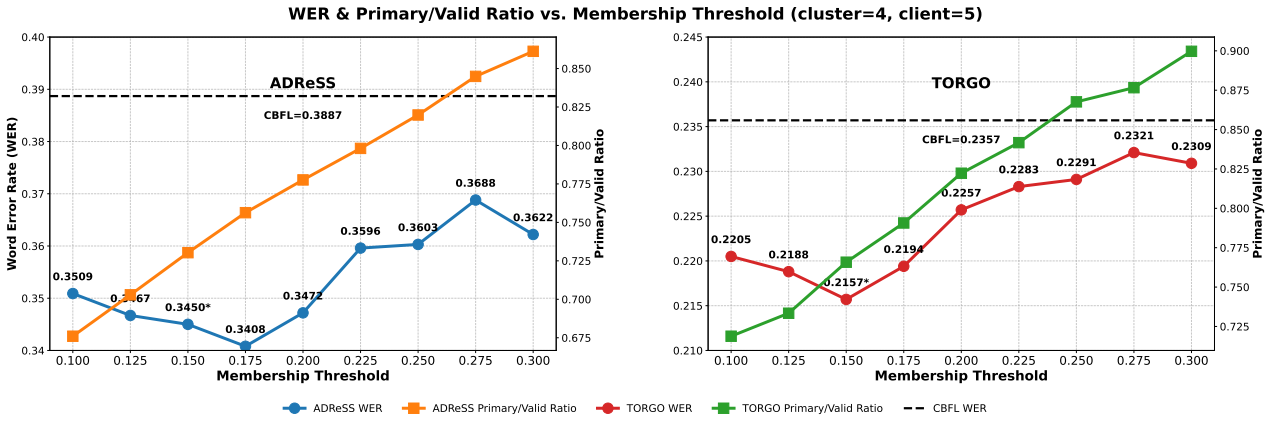

Fig. 3. Membership threshold analysis showing the relationship between Primary/Valid Ratio and WER performance in FCPFL.

demonstrates medium-to-large effect size over CPFL-embs (t = 2.64, p $<$ 0.01, Cohen's d = 0.64). Bootstrap validation (1,000 iterations) provides robust 95% confidence intervals supporting these gains. Table II presents key metrics quantifying FCPFL's robustness advantages. FCPFL reduces performance variance by 61.5% (ADReSS) and 29.5% (TORGO), with strong correlations between sample availability and WER (ADReSS: $r = 0.911$, $p = 0.002$; TORGO: $r = 0.810$, $p = 0.015$), demonstrating predictable behavior that hard clustering lacks.

VI. Conclusion

In this work, we have introduced Fuzzy Cluster-Based Personalized Federated Learning (FCPFL) to address the challenges of privacy preservation and extreme heterogeneity in pathological speech recognition. By leveraging Fuzzy C-Means to compute soft membership degrees and incorporating pseudo-label–guided feature selection, FCPFL enables samples near cluster boundaries to participate in multiple clusters, thus increasing the effective training set and improving model generalization under non-IID conditions.

Comprehensive experiments on both ADReSS and TORGO datasets demonstrate that FCPFL consistently outperforms hard-clustering and standard federated baselines, yielding relative WER reductions of 4.82% on ADReSS and 1.74% on TORGO. Our hyperparameter analyses reveal four key insights: both cluster number and threshold variations affect performance through the unified mechanism of boundary sample utilization; FCPFL maintains consistent robustness across parameter dimensions due to its fuzzy membership design; the strong correlations demonstrate statistically reliable and predictable behavior; and this robustness to hyperparameter variations makes FCPFL practically applicable for real-world deployment where optimal parameters are unknown a priori.

Beyond optimizing fuzzy clustering parameters, we observed that boundary samples exhibit different distributions across datasets, which may carry implicit information about underlying disease features. These boundary sample distributions could potentially assist in defining disease classifications and characteristics, providing new insights into pathological speech patterns. In future work, we will collect more diverse pathological speech data to investigate whether these boundary-sample distributions can serve as biomarkers for disease progression or subtype, potentially enabling more targeted clinical insights and diagnosis.

REFERENCES

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.

[2] M. A. Little and P. A. Foltynie, "A review of dysarthric speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, no. 5, pp. 975–988, 2021.

[3] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[4] C.-L. Hsu, D. Smith, and T.-Y. Chen, "Clustered personalized federated learning for dysarthric speech recognition," in *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, 2023, pp. 1–13.

[5] X. Wang, J. Liu, and K. Zhao, "Conformer models for elderly speech recognition: Robustness to acoustic variability," in *Proc. Interspeech*, 2022, pp. 1450–1454.

[6] D. Wang, W. Zhang, B. P. Lim, J. Guo, and H. Meng, "Recent progress in the CUHK dysarthric speech recognition system," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2267–2281, 2021.

[7] O. El-Rifai, J. Nguyen, and S. Patel, "Clustered federated learning: A survey of methods and applications in healthcare," *IEEE Trans. Biomed. Eng.*, vol. 72, no. 4, pp. 990–1008, 2025.

[8] S. Mintz, R. Kumar, and L. Zhang, "Challenges of non-IID data in federated learning: A comprehensive survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3504–3520, 2022.

[9] C. Dwork and A. Roth, *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science, 2014, vol. 9, no. 3–4.

[10] D. Smith, Y. Li, and C. Brown, "Personalized federated learning: Handling heterogeneity and uncertainty," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 2, pp. 234–248, 2023.

[11] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, and et al., "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[12] D. Guliani, F. Beaufays, and G. Motta, "Joint federated learning and personalization for on-device ASR," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 966–10 970.

[13] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "FedProx: A federated optimization strategy with proximal terms," in *Proc. 23rd Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2020, pp. 1308–1318.

[14] Y. Zhao, M. Li, L. Lai, and N. Suda, "Federated learning with non-IID data: Analysis and solutions," in *ICML Workshop on Federated Learning*, 2018.

[15] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, 2020.

[16] ——, "Federated optimization in heterogeneous networks," in *Proc. Third MLSys Conference*. PMLR, 2020, pp. 1–16.

[17] A. Chen, Y. Zhou, and K. Li, "FedMeta: Federated meta-learning for personalized recommendations," in *Proc. 27th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, 2021, pp. 2500–2508.

[18] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. 37th Int. Conf. Machine Learning (ICML)*. PMLR, 2020, pp. 699–710.

[19] Y. Huang, X. Zhang, and J. Lin, "Patient clustering for federated healthcare analytics," *IEEE Trans. Big Data*, vol. 5, no. 3, pp. 456–467, 2019.

[20] Q. Li, B. Xiong, X. Zhang, F. Shen, and D. Yin, "Performance of federated learning for medical image analysis in realistic scenarios: A case study in brain tumor segmentation," *Medical Image Analysis*, vol. 67, p. 101835, 2021.

[21] 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, pp. 2172–2176.

[22] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2–3, pp. 191–203, 1984.

[23] W. Hu, Q. Liu, and Z. Guo, "Federated multi-view clustering with fuzzy c-means," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 1124–1137, 2023.

[24] J. Liu, Y. Wu, Q. Jiang, and H. Zhang, "Deep fuzzy c-means clustering for high-dimensional data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3060–3073, 2021.

[25] S. Park, H. Kim, and J. Seo, "Pseudo-labeling in semi-supervised speech recognition," in *Proc. Interspeech*, 2021, pp. 3120–3124.

[26] J. Lee, M. Chen, and T. Huang, "Low-resource speech recognition via pseudo-label generation," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing (ICASSP)*, 2020, pp. 7899–7903.

[27] K. Sohn, D. Berthelot, C. Li, Z. Zhang, N. Carlini, E. Cubuk, A. Kurakin, and F. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 596–608.

[28] A. Sapkota, R. Thapa, and L. Zhao, "Dysarthric speech feature ranking and federated learning," in *Proc. IEEE Int. Conf. Healthcare Informatics (ICHI)*, 2025, pp. 45–52.

[29] F. Rudzicz, A. K. Amavasya, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 523–541, 2012.

[30] J. Liu, A. Kumar, M. Yu, and N. Sharma, "Federated representation learning for automatic speech recognition," *Amazon Science*, 2023.

[31] H. Mehmood, A. Dobrowolska, K. Saravanan, and M. Ozay, "FedNST: Federated noisy student training for automatic speech recognition," in *ICML*, 2022.

[32] Y. Du, Z. Zhang, L. Yue, X. Huang, Y. Zhang, T. Xu, L. Xu, and E. Chen, "Communication-efficient personalized federated learning for speech-to-text tasks," 2024, arXiv:2401.10070.

[33] Z. Li, K. Zhao, J. Zhou, and Y. Xiao, "FedMP: A multi-cluster federated learning framework for personalized models," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 34, no. 04. AAAI Press, 2020, pp. 5100–5107.

[34] M. Stallmann and A. Wilbik, "On a framework for federated cluster analysis," *Applied Sciences*, vol. 12, no. 20, p. 10455, Oct. 2022.

[35] X. Hu, J. Qin, Y. Shen, W. Pedrycz *et al.*, "Federated fuzzy clustering for decentralized incomplete longitudinal behavioral data," 2023, arXiv:2308.00000.

[36] C. Wang, W. Pedrycz, Z. Li, M. Zhou, and J. Zhao, "Residual-sparse fuzzy c-means clustering incorporating morphological reconstruction and wavelet frame," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 12, pp. 3910–3921, 2021.

[37] Z. Zhang, A. Smith, and L. Chen, "SplitAVG: Heterogeneity-aware federated learning in medical imaging," in *Proc. Int. Conf. Medical Image Computing Computer-Assisted Intervention (MICCAI)*, 2021, pp. 234–242.

[38] J. Liu, A. Kumar, M. Yu, and N. Sharma, "Federated representation learning for automatic speech recognition," *Amazon Science*, vol. 5, pp. 123–134, 2022.

[39] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841–847, 1991.

[40] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[41] 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 *Proc. Int. Conf. Machine Learning (ICML)*, vol. 162. PMLR, 2022, pp. 1298–1312.

[42] 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.

[43] F. Rudzicz, G. Hirst, P. van Lieshout, G. Penn, F. Shein, A. Namasivayam, and T. Wolff, "TORGO database of dysarthric articulation," Web Download. Linguistic Data Consortium, LDC2012S02, 2012.

[44] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Machine Learning and Systems*, vol. 2, 2020, pp. 429–450.

[45] S. Ping, L. Wang, L. Nie, Y. Ma, and C. Xu, "FL-TAC: Enhanced fine-tuning in federated learning via low-rank, task-specific adapter clustering," 2024, arXiv:2404.15384.