RESEARCH

Learning Domain-Heterogeneous Speaker Recognition Systems with Personalized Continual Federated Learning

Zhiyong Chen[1]
and Shugong Xu[2*]

Abstract

Speaker recognition, the process of automatically identifying a speaker based on individual characteristics in speech signals, presents significant challenges when addressing heterogeneous-domain conditions. Federated learning, a recent development in machine learning methods, has gained traction in privacy-sensitive tasks, such as personal voice assistants in home environments. However, its application in heterogeneous multi-domain scenarios for enhancing system customization remains underexplored. In this paper, we propose the utilization of federated learning in heterogeneous situations to enable adaptation across multiple domains. We also introduce a personalized federated learning algorithm designed to effectively leverage limited domain data, resulting in improved learning outcomes. Furthermore, we present a strategy for implementing the federated learning algorithm in practical, real-world continual learning scenarios, demonstrating promising results. The proposed federated learning method exhibits superior performance across a range of synthesized complex conditions and continual learning settings, compared to conventional training methods.

Keywords: speaker recognition; federated learning; domain adaptation; continual learning

Introduction

Speaker recognition, a critical task in the field of speech processing, involves the automatic identification and verification of speakers based on individual characteristics embedded in speech signals. With the growing ubiquity of voice-controlled devices and systems, speaker recognition has become an essential component for various applications, including security, authentication, and personalized user experiences.

Deep neural networks have become the cornerstone of modern machine learning applications, often requiring large amounts of labeled training data to achieve optimal performance. Traditionally, this data is collected from end-devices, such as smartphones, and sent to a centralized server for model training. However, this approach raises concerns regarding user privacy and the potential burden on communication links due to the transmission of large datasets.

In recent years, researchers in speaker recognition field are putting more focus on learning robust speaker features on multiple conditions [1, 2]. Including different room acoustics scenarios, different languages, different channel conditions, etc. All these contribute to degraded speaker recognition performance. Many researches focus on using domain adaptation methods to improve the system performance in these scenarios. While many of these research need to obtain both target domain data and the source domain data in a central data center, which is not only cost inefficient, but also sometimes impossible.

Federated Learning (FL), an emerging machine learning paradigm, has gained significant attention in recent years for its potential to improve privacy and enable collaborative learning among distributed data sources. FL allows multiple clients to jointly train a model without sharing raw data, which can be particularly useful in privacy-sensitive applications. Despite its growing popularity, the application of FL in heterogeneous multi-domain conditions for enhancing system customization in speaker recognition remains relatively unexplored. Federated learning can be broadly categorized into two main types [3]:

- Cross-device Federated Learning: This method focuses on jointly learning speech characteristics

*Correspondence: shugong@shu.edu.cn
[2]School of Communication and Information Engineering, Shanghai University, Shanghai, China
Full list of author information is available at the end of the article

from numerous mobile or similar devices to train a unified statistical model for speaker recognition. In this typical scenario, data is often limited and may have lower labeling quality.
- Cross-silo Federated Learning: In this setting, organizations like universities can be regarded as remote devices containing substantial student data. These organizations must adhere to strict privacy practices and navigate potential legal, administrative, or ethical constraints to ensure data privacy. Federated learning can be employed in this scenario with relatively more abundant data and better labeling quality, facilitating the construction of supervised yet cost-effective training in this situation.

Privacy concerns are regarded as one of the major challenges in speaker recognition applications since they involve the complete sharing of speech data, which can have serious implications for user privacy. Federated learning can mitigate privacy infringement in speaker recognition systems by enabling multiple participants to collaboratively learn a shared model without revealing their local data, as recently examined by Lai et al. [4]. Interestingly, an emerging trend in the FL domain involves utilizing federated learning for domain adaptation and personalization, leading to a research area known as personalized FL [5]. This approach eliminates the need for centralized data transmission, storage, and training, making adaptation to diverse and complex client conditions more feasible and reasonable. Another challenge in real-world scenarios includes ever-changing data with limited buffer capacity, prompting research into continual learning and online learning [6].

The main contribution of this work is the application of federated learning techniques to train supervised deep neural network-based speaker recognition models, with the goal of customizing speaker information across multiple heterogeneous domains while preserving user privacy. Unlike previous works on FL in combination with speech and speaker recognition, which mainly focus on privacy-preservation scenarios and simple client collaboration within a single data domain, we concentrate on multi-domain client collaboration and heterogeneous domain adaptation using personalized FL. To achieve this, we simulate iconic acoustic conditions using the room acoustic software Pyroom [7] and select multi-lingual datasets to design and compose client datasets. We also evaluate various personalized training strategies to identify a better approach that outperforms centralized training. Finally, we explore ways to combine FL methods with continual learning techniques, enabling them to function effectively in real-world continual learning scenarios.

In summary, this paper explores the effectiveness of Personalized Federated Learning (PFL) and Federated Learning combined with continual learning methods within the scope of speaker identification and speaker verification tasks across multiple heterogeneous domains. Our primary contributions can be outlined as follows:
- We propose a speaker recognition system based on Personalized Federated Learning (PFL), leveraging supervised speaker data stored across different silos. By learning client-dependent projection modules, our approach enables better adaptation to various scenarios and demonstrates promising performance in both speaker identification and speaker verification tasks.
- We simulate and evaluate our systems using room acoustics software to assess PFL's effectiveness in domain adaptation scenarios. We compare PFL's performance with centralized training and other common baselines, showing that PFL surpasses these alternatives, a finding unreported in other FL-based speech research. Our carefully designed training strategies demonstrate that the proposed PFL methods are particularly suitable for domain-heterogeneous speaker recognition scenarios.
- We effectively integrate federated learning with continual learning settings, introducing Continual Personalized Federated Learning (C-PFL) that delivers robust performance throughout training stages. Our chosen random prototype casting training strategy, employed as an enhancement, proves to be beneficial when combined with C-PFL.

The remainder of this paper is organized as follows: Section II reviews related work. Section III describes the proposed model. Section IV details the experimental setup and presents the results and analysis. Finally, Section V concludes the paper.

Related Work

Speaker Recognition

Automatic speaker recognition has a rich history, with early methods including probabilistic models, deep neural networks combined with probabilistic models, and end-to-end speaker recognition models [8, 9, 10, 11]. Speaker recognition encompasses three subtasks: speaker verification, speaker identification [1], and speaker diarization [12]. This paper primarily focuses on speaker verification and speaker identification tasks.

Over the past decade, neural network-based speaker recognition models have achieved superior performance, becoming the dominant approach in the field.

The x-vector model, which extracts speaker-related features from acoustic properties using neural networks [9], can be considered a milestone in modern deep neural network speaker recognition. Subsequent years have seen the development of convolutional-based [13], complex 1D temporal neural network-based [10], transformer-based DNN systems [14], and autoML-based systems [15], all of which have contributed to significant progress in speaker recognition.

Following the trend of large-scale training, speaker recognition models that leverage transformer-based pre-training [16, 17] have demonstrated impressive improvements over previous deep learning methods.

Domain adaptation & robust speaker recognition

Although speaker recognition systems have demonstrated strong performance on numerous benchmark datasets, recognizing speech in complex and diverse domains remains a challenging problem. Current methods for addressing domain adaptation issues can be categorized into several groups. Back-end statistical model adaptation techniques [18, 19] utilize first- and second-order information from the embedding space feature distribution to adapt the backend classification model. These models are generally lightweight [20], resulting in high explainability.

Recently, many works have focused on developing methods that can better learn from different datasets or conditions, employing transfer learning [21] approaches such as domain adversarial learning [22, 23] or discrepancy minimization methods [24]. Other techniques concentrate on using meta-learning methods to construct domain-agnostic pretrained models from various datasets [25] or adapting parts of the base pretrained model to build better target domain representations [26].

In order to enhance the robustness of speaker recognition models, some researchers utilize data augmentation methods. Notable recent works include using text-to-speech techniques to synthesize fake speakers [27, 28], which helps make the speaker model more generalizable. Other approaches involve more sophisticated audio signal processing technologies in the front-end, such as beamforming methods [29], dereverberation techniques [30], and speech separation methods [31], all of which contribute to making speaker recognition systems more robust in varying acoustic conditions.

Continual learning & its application

Continual learning and its related online learning scenarios and methods are gaining more attention recently [6]. In recent research, several approaches have been proposed to tackle the challenges of continual learning and generalization in various domains, including speaker verification and automatic speech recognition. In [32], the authors propose a continual-learning-based method to incrementally learn new spoofing attacks for speaker verification systems without performance degradation on previous data. Paper [33] presents a dynamically expanding end-to-end model for the speech recognition task, which helps avoid catastrophic forgetting and seamlessly integrate knowledge from new data. Paper [34] focuses on online continual learning for automatic speech recognition and demonstrates the effectiveness of incremental model updates using the online Gradient Episodic Memory (GEM) method.

Federated Learning & its application

Federated learning has emerged as a promising technology in the field of machine learning, with significant potential for preserving user privacy [35]. In the speech community, numerous studies have employed federated learning techniques in various applications such as automatic speech recognition [36, 37, 38, 39, 40], keyword spotting [41, 42], and speech emotion detection [43].

A number of works have also applied federated learning methods to speaker recognition tasks, including [44, 4, 45]. These studies primarily focus on utilizing federated learning to enhance data privacy and explore the data class distribution properties of each client in non-IID scenarios within the same domain condition.

Learning Speaker Features with Personalized Federated Learning

The Federated Learning-based speaker recognition system consists of two learning procedures: client-side fine-tuning and server-side updates. Federated Learning allows for distributed training of speaker recognition models across domain-heterogeneous clients. Our proposed Personalized Federated Learning (PFL) system for speaker recognition operates in two primary locations: edge silos (clients) and a central server. The overall architecture of the system, showcasing the server-side, silos' domain conditions, and individual client learning details, is illustrated in Figure 1.

$$\min_{\theta} G(\theta), \text{ where } G(\theta) := \sum_{i=1}^{n} q_i G_i(\theta) \tag{1}$$

In Equation (1), we aim to minimize the global objective function $G(\theta)$, which is defined as the weighted sum of local objective functions $G_i(\theta)$ across n clients. Here, q_i denotes the weights for aggregating the targets, and θ represents the model parameters.

This formulation illustrates the process of training a centralized model over a distributed dataset, where a

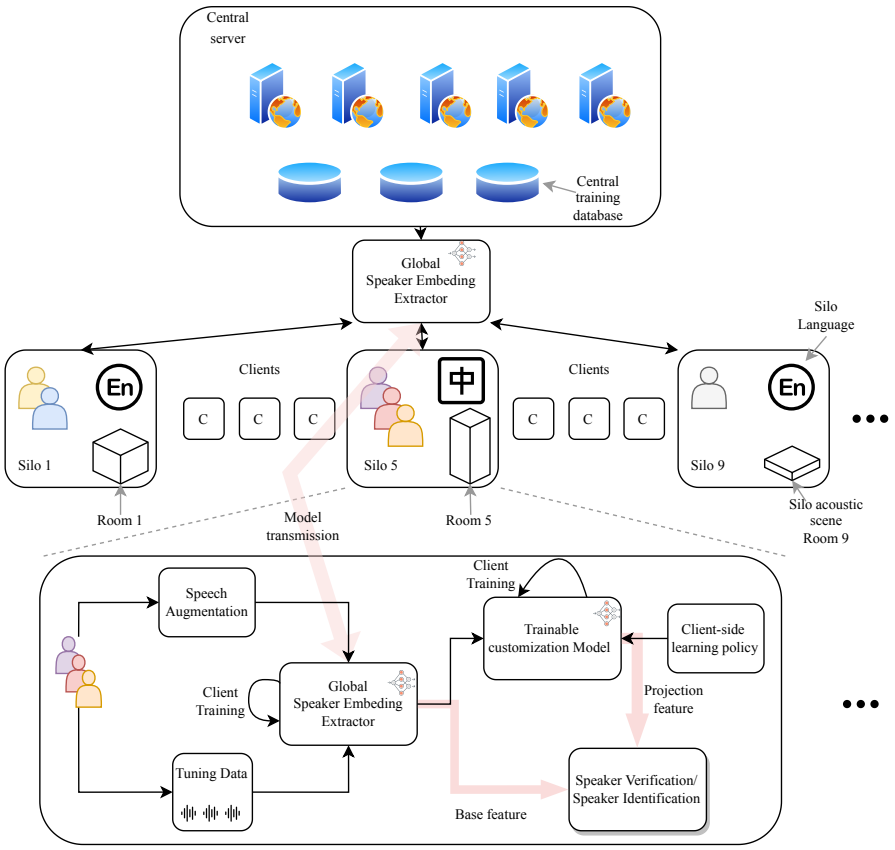

Figure 1 Architecture of the proposed multi-domain personalized federated speaker system. The proposed architecture is used in major speaker recognition sub-tasks including speaker verification and identification.

multitude of clients hold variable-sized subsets of the data. During each iteration of training, a local model update is computed at the device level and communicated to a central server. Subsequently, the central server combines a large number of these updates or gradients to compute a global update to the central model. This global update is essentially an average of the local updates, which ensures the preservation of privacy and efficient utilization of distributed data.

$$G(\theta, X, Y) := \text{LossFunc}(\text{Trans}(\text{Base}(X)), Y) \quad (2)$$

Equation (2) demonstrates the core concept of personalized FL, where Base$(\cdot)$ serves as the model combination, employing a global feature extractor based on x-vectors in our setup. Meanwhile, Trans$(\cdot)$ represents a Transformer-based personalized feature projector for each domain-specific client before calculating the final loss for each client. This approach adopts the Transfer Learning-based personalized FL (TL-PFL) strategy, as described in [5].

Given an speaker embedding from the Base$(\cdot)$ and sequenlize it as $X = \{x_1, x_2, \ldots, x_n\}$, the Transformer encoder first applies a positional encoding to the input embeddings, which allows the model to utilize positional information. The encoded sequence $Z^0 = \{z_1^0, z_2^0, \ldots, z_n^0\}$ is then fed into the first layer of the encoder. Each layer l computes the output sequence $Z^l = \{z_1^l, z_2^l, \ldots, z_n^l\}$.

In the Transformer encoder, the Multi-Head Attention mechanism is denoted as MultiHead(Q, K, V), where Q, K, and V represent the query, key, and value matrices, respectively, serving as the inputs for this mechanism.

For each layer of the Transformer encoder:

$$Z^{l,\text{att}} = \text{MultiHead}(Z^{l-1}, Z^{l-1}, Z^{l-1}) \quad (3)$$

$$Z^{l,\text{ff}} = \text{FFN}(Z^{l,\text{att}}) \quad (4)$$

$$Z^{l} = \text{LayerNorm}(Z^{l-1} + Z^{l,\text{att}}) \quad (5)$$

$$Z^{l+1} = \text{LayerNorm}(Z^{l} + Z^{l,\text{ff}}) \quad (6)$$

Algorithm 1 Server update algorithm for PFL & C-PFL.

Initialize w^0 as pre-trained weights
for all epoch $e = 1, 2, ...$ **do**
  $m \leftarrow$ (select random subset m of M clients, each with probability P)
  for all m_i clients in m **do**
    $\hat{w}^e_{m_i} \leftarrow$ ClientUpdate (m_i, w^e) **Algorithm 2**
    $\Delta w^e_{m_i} = w^e - \hat{w}^e_{m_i}$
  end for
  $n_{m_i} \leftarrow$ CountDataForEachClient ()
  $n = \sum_{m_i=1}^{m} n_{m_i}$
  $\bar{w}^e = \sum_{m_i=1}^{m} \frac{n_{m_i}}{n} \Delta w^e_{m_i}$
end for
$w^{e+1} = w^e - \eta \bar{w}^e$

Algorithm 2 Client update algorithm in C-PFL.

Input: CL Stage: t, Client's local datset: D_{local}, Client's ID: m_i
Initialize:
Set learnable speaker prototypes in **W** as defined in (8)(13)(14):
$c_t \leftarrow$ RandomInitializeLabel (t, D_{local})
During Fedetered Learning:
Receive w^e from the server from **Algorithm 1**
for all local epoch $e_{local} = 1, 2, ...$ **do**
  $D_e, y_e \leftarrow$ SamplingLocalDataset (D_{local}, c_t)
  $G \leftarrow$ Forwarding(D_e, y_e) using Equation (2)
  $\hat{w}^e_{m_i} \leftarrow$ GradientUpdate (G)
  $\Delta w^e_{m_i} = w^e - \hat{w}^e_c$
end for
return Δw^e_c

Where FFN represents the position-wise feed-forward network, and LayerNorm denotes layer normalization.

As illustrated in Figure 1, the final output embedding for evaluation is produced by two branches, namely the Base Feature (referred to as Type-A, Discriminator-projection) and Projection Feature (referred to as Type-B, Feature-projection), which can be used individually or in combination. The embedding is used directly for cosine similarity comparison in speaker verification tasks following [10]. For speaker identification tasks, a separate logistic regression-based classifier is learned for each speaker in each evaluation subset:

$$P(c_{id}|\mathbf{x_{emb}}; \mathbf{w}) = \sigma(\mathbf{W_{id}^{(i)}}^T \mathbf{x_{emb}}) \tag{7}$$

Here, c_{id} represents the identification output class ID, $\mathbf{W_{id}^{(i)}}^T$ denotes the learnable weight for the i-th evaluation subset, and $\mathbf{x_{emb}}$ is the generated speaker embedding.

Heterogenous Domain Continual Personalized Federated Learning

We present the Heterogeneous-domain Across-silo Continual Personalized Federated Learning (C-PFL) method, which combines the principles of continual learning and federated learning. Our approach is specifically designed to address the challenges encountered when new data from different stages become available, all while preserving the aforementioned FL mechanism across diverse silos.

The key idea behind our method is to dynamically update the output classifier parameters with new weights when data from new stages arrive. This is achieved by continually adapting the model to incorporate the information from the newly acquired data without causing significant interference with the previously learned knowledge.

In the proposed C-PFL method, given the output of speaker embedding as $\mathbf{x} \in R^n$, the output probability for each class is estimated by the equation:

$$P(c|\mathbf{x}; \mathbf{w}, \mathbf{b}) = \sigma(\mathbf{W}^T\mathbf{x} + \mathbf{b}) \tag{8}$$

Where $c \in N_0$ is the non-negative label ID for the speaker embedding, $\sigma(\cdot)$ is the Softmax function, and $\mathbf{W} \in R^{C\times n}$ and $\mathbf{b} \in R^C$. Given that we have T continual learning stages in which each stage has a speaker set $\mathbf{c_t}$, we configure C as follows:

$$\Sigma_{t=1}^{T}|\mathbf{c_t}| \ll C \tag{9}$$
$$c_{ti} \sim \text{Uniform}(0, C) \tag{10}$$

Where c_{ti} is the assigned label for speaker embedding i in the t-th continual learning stage. As new data arrive at different stages, the output classifier sets new weights specifically for the classes in these new stages with high probability if C is set large enough:

$$P_{new}(t) > 1 - \frac{\Sigma_1^{n-1}|\mathbf{c_t}|}{C} \tag{11}$$
$$\lim_{C\to\infty} P_{new}(t, C) = 1 \tag{12}$$

Here, $P_{new}(t)$ is the probability of setting new classes at the continual learning stage t. Let $\Omega^{(t)}$ denote the classifier weights vector sets (also called prototypes) in Equation (8) used at stage t, which is defined as:

$$\Omega^{(t)} = \{W_{:,c_{ti}}|c_{ti} \in \mathbf{c_t}\} \tag{13}$$
$$\Omega^{(t_1)} \cap \Omega^{(t_2)} = \emptyset; (t_1 \neq t_2) \tag{14}$$

Equation (14) holds true when the condition in (9) is met. In the initial learning stage, we employ weights $\Omega^{(1)}$. As new data becomes available for subsequent stages, we introduce additional weights $\Omega^{(2)}$, and so forth. This method enables the model to learn continually from new data without experiencing catastrophic

forgetting of previously acquired knowledge. The detailed procedure for PFL and C-PFL are shown in Algorithm 1 and Algorithm 2. This approach, named "C-PFL with enhanced training strategy," is contrasted with a simpler method that employs Equation (8) without random prototype casting, which we designate as "C-PFL with a standard training strategy," to emphasize the differences between the two strategies.

Silo Acoustics Simulation

To evaluate the PFL and C-PFL methods in scenarios involving multiple domain clients, we simulated room acoustic settings in our experiments using the Pyroom Acoustics library. Six distinct room configurations were employed to thoroughly examine the performance of our models under challenging acoustic conditions. As illustrated in the Figure 2, these room settings encompassed a diverse range of sizes, shapes, and reverberation characteristics, thereby providing a comprehensive evaluation framework. In each room, we tested the models using both single OMNI-directional microphones and OMNI-directional microphone arrays. This approach enabled us to investigate the impact of different microphone configurations on the overall system performance, as well as assess the robustness of our models in coping with diverse real-world scenarios.

The first three rooms are designed with varying sizes to simulate different room reverberation properties, representing small, medium, and large rooms. The left part of Figure 3 demonstrates the RT60 metrics of each room, showcasing their distinct reverberation characteristics. The fourth room includes a noise source alongside the sound source to simulate noisy room conditions.

Our room simulations also incorporate delay-and-sum (DAS) beamforming to emulate the performance of real-world multi-channel microphone array acoustic environments and evaluate how our systems perform under these common conditions. The right part of Figure 3 depicts the beamforming settings for Room5 and Room6, presenting the beam patterns across different frequency bands. Room6 additionally includes a noise source, as in Room4. The delay-and-sum beamforming algorithm is applied in these two acoustic scenarios. Let $x_i(t)$ denote the signal received by the i^{th} microphone in these rooms, where $i = 1, \ldots, M$, and M is the total number of microphones. The time-delay τ_i is calculated for each microphone based on the desired direction of the beam, as expressed below:

$$\tau_i = \frac{d_i}{c} \tag{15}$$

$$y(t) = \sum_{i=1}^{M} x_i(t - \tau_i) \tag{16}$$

where d_i represents the difference in distance between the source and the i^{th} microphone and the reference microphone, and c is the speed of sound. The output signal $y(t)$ is then computed by summing the delayed signals from all microphones to enhance the preset direction.

Experiments

Experiments Settings

We utilize simulated room acoustics scenarios to assess the federated learning system and its associated algorithms in the context of speaker recognition tasks. The evaluation specifically targets both speaker verification and speaker identification tasks across multiple heterogeneous domain groups, and we also investigate the performance in continual learning settings. The configurations can be found in Table 1.

To build the speaker recognition system with personalized federated learning, we use the VoxCeleb [46] and CnCeleb [47] datasets. We select 100 speakers for each group, ensuring no overlap between groups. Each group simulates domain conditions based on the predefined settings described in the table, enabling a comprehensive assessment of the system's performance across various acoustic environments and languages. We configure Groups 1 through 6 to use the settings of Rooms 1 through 6 with the English language VoxCeleb dataset, while Groups 7 through 12 employ the same room settings but with the Chinese language CnCeleb dataset.

For system evaluation, speaker IDs are selected according to the table for both verification and identification tasks. The evaluation speakers for each group remain consistent across all groups, with their speech processed by the simulation pipeline according to the predetermined settings.

In the continual learning settings, we additionally allocate Stage 2 and Stage 3 data using distinct speakers from VoxCeleb, while keeping the evaluation set consistent across all stages to evaluate the performance of the proposed systems during each stage.

Table 1 also documents detailed settings for each group, encompassing room dimensions, reverberation time, microphone array configurations, and other relevant parameters.

We utilize the VoxCeleb2 dataset to pre-train the x-vector model for the canonical model, following the

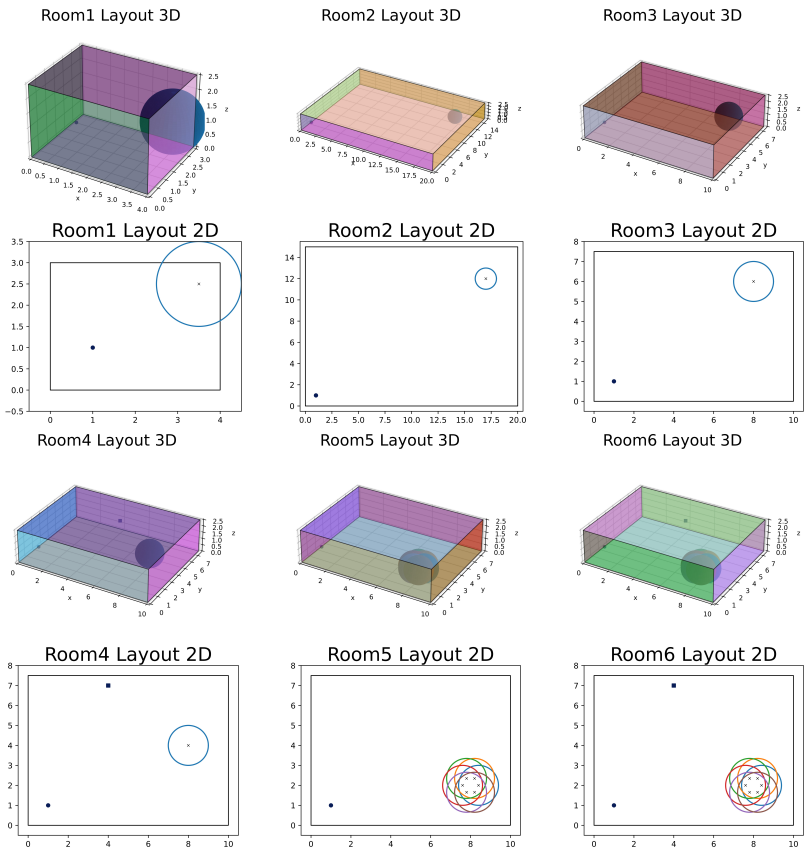

Figure 2 The room acoustic settings simulated with Pyroom acoustics. Six rooms settings used in the experiments are demostrated in the figure. Rooms are using both single OMNI-direction microphone or OMNI-direction mic-arrays.

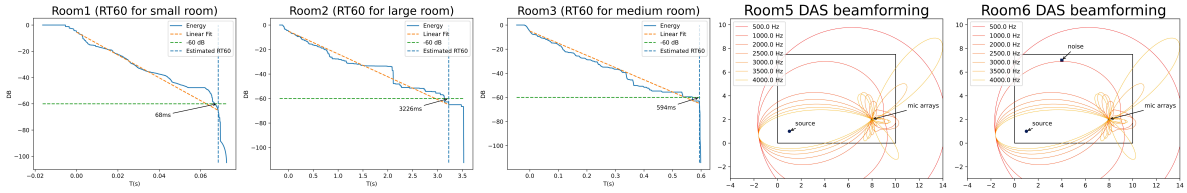

Figure 3 Detailed room simulation information: The first three figures display the RT60 properties of the small, large, and medium rooms in our simulation. The last two figures illustrate the beamforming settings in Room5 and Room6, as well as the beam patterns across different frequency bands.

same procedure as in [10]. Subsequently, we conduct further experiments with PFL, C-PFL, and other baseline algorithms based on this pre-trained model.

Performance Evaluation of Personalized Federated Learning

In Table 2, we present the evaluation results for various models across evaluation groups 1 to 12, along with their mean values. Our findings indicate that the Canonical model, following the classical training procedure of [10], demonstrates satisfactory performance in certain domain groups that exhibit similar acoustic conditions to the original dataset. However, its performance is limited in many other groups characterized by distinct acoustic conditions.

The Centralized training model, incorporating a transformer block similar to the architecture of [14], exhibits a substantial improvement over the Canonical models, as it leverages the domain information collec-

Table 1 Dataset settings for experiments

Group	Scenario settings (Room/Language)	FL training ID (Base-S1/CL-S2/CL-S3)	Speaker ID for eval
G1	Room1/EN	Vox1 1-100/601-650/901-950	Vox1 Test 1-40
G2	Room2/EN	Vox1 101-200/651-700/951-1000	
G3	Room3/EN	Vox1 201-300/701-750/1001-1050	
G4	Room4/EN	Vox1 301-400/751-800/1051-1100	
G5	Room5/EN	Vox1 401-500/801-850/1101-1150	
G6	Room6/EN	Vox1 501-600/851-900/1151-1200	
G7	Room1/CN	CnCelb 1-60/-/-	CnCelb 331-430
G8	Room2/CN	CnCelb 61-120/-/-	
G9	Room3/CN	CnCelb 121-180/-/-	
G10	Room4/CN	CnCelb 181-240/-/-	
G11	Room5/CN	CnCelb 241-300/-/-	
G12	Room6/CN	CnCelb 301-330/-/-	

Table 2 Evaluation performance of PFL and the baseline

Evaluation Groups	Canonical model [9][10] *EER*	Centeralized training [14] (Assembling fine-tuning) *EER*	Seperated training [26] (Individual fine-tuning) *EER*	PFL (Ours) Type-A *EER*
G1	**2.85**	9.664	13.712	4.978
G2	33.06	17.333	31.022	**14.36**
G3	**4.81**	10.627	14.174	6.864
G4	39.23	21.189	27.652	**19.961**
G5	**5.51**	10.351	17.977	6.274
G6	37.80	18.679	27.662	**17.168**
G7	16.25	18.86	24.728	**15.412**
G8	41.29	24.032	38.723	**24.324**
G9	20.61	18.975	30.294	**16.554**
G10	37.61	25.046	34.309	**23.73**
G11	24.45	19.419	33.983	**16.779**
G12	37.10	24.071	33.149	**22.336**
mean(μ)	25.0475	17.44	28.87	**15.01**

Table 3 Comparing performance of different training strategies of PFL

Evaluation Groups	FL Classical [4] Direct-discriminator *EER*	PFL Type-A Discriminator-projection *EER*	PFL Type-B Feature-projection *EER*
G1	5.297	**5.037**	5.481
G2	15.089	**14.100**	25.242
G3	**6.935**	**6.975**	7.239
G4	19.961	**18.986**	20.485
G5	**6.375**	6.680	7.438
G6	17.355	**15.796**	19.703
G7	**15.799**	16.012	16.902
G8	**24.619**	25.096	31.103
G9	**17.153**	17.443	20.49
G10	23.805	**23.653**	25.927
G11	**17.448**	17.871	21.632
G12	**22.806**	22.993	26.048
mean(μ)	16.053	**15.887**	19.137

tively within the central server. In contrast, the Separated training procedure, employing a strategy similar to [26], does not consistently yield better performance than the Canonical model. This highlights the impact of limited data for fine-tuning, which may lead to degraded performance compared to the original models.

The Personalized Federated Learning strategy outperforms all other methods in these domain-agnostic scenarios, achieving the lowest mean EER among all approaches. The improvement is particularly significant in groups with rare acoustic conditions that deviate considerably from the original training data. This finding underscores the effectiveness of the PFL strategy in addressing speaker recognition tasks across diverse acoustic environments.

In Table 3, we evaluate various Personalized Federated Learning (PFL) strategies, comparing the classical direct-discriminator approach, as used in [4], to the proposed personalized projection-based methods. Both methods yield promising results, with PFL achieving a better mean EER using the discriminator-projection strategy (PFL Type-A). Therefore, we further assess these two techniques by examining their convergence capabilities, as illustrated in Figure 5. Uti-

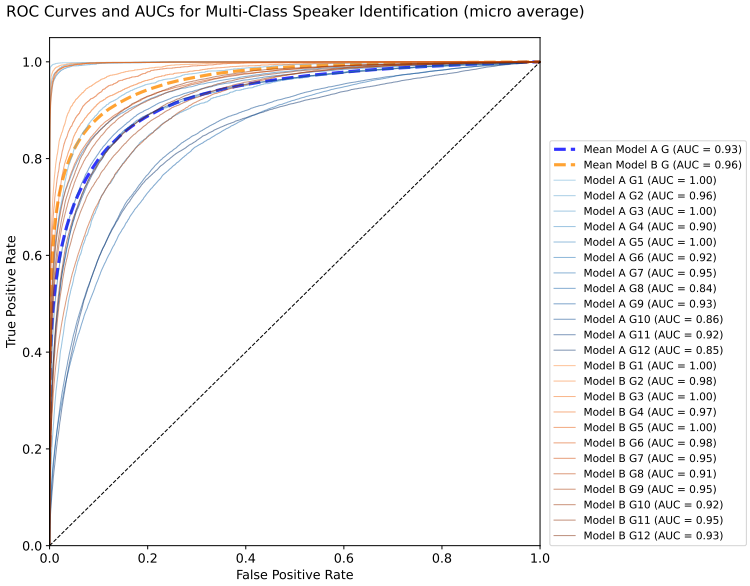

Figure 4 Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values for multi-class speaker identification using Models A (Centralized training) and B (PFL Type-A). The plot shows individual ROC curves for each of the 12 groups, with different shades of blue for Model A and different shades of orange for Model B. The mean ROC curves for Model A and Model B are highlighted with thicker dashed lines. The AUC values for each curve are included in the legend.

lizing a personalized discriminator demonstrates significantly better convergence performance, achieving the lowest EER within just 10 rounds, whereas the direct-discriminator approach requires approximately 25 rounds. This showcases the efficiency and effectiveness of the PFL method.

Moreover, we investigate the effects of various FL strategies on overall performance. Utilizing a feature-projection strategy (PFL Type-B) does not lead to better outcomes, indicating that the limited data available for fine-tuning the transformer-based projector presents challenges in creating a more efficient feature space. These insights emphasize the strengths and limitations of different Personalized FL strategies when addressing speaker recognition tasks in diverse acoustic environments.

Speaker Identification Task Evaluation

We further evaluate the performance of the Personalized-FL (PFL) model by comparing its performance on the speaker identification task. The results are presented in Table 4. We compare various basic classification metrics across the 12 evaluation domain sets, including accuracy, precision, F1-score, and AUC. The corresponding micro-average ROC curves for each domain group are also assessed, as depicted in Figure 4.

Our analysis reveals that using our proposed PFL strategy leads to significantly better results than centralized training, with the exception of some domain groups that have relatively more common and less challenging acoustic conditions. In these cases, both methods perform similarly. The mean ROC curve in Figure 4 further illustrates this trend, providing a comprehensive visualization of the PFL strategy's effectiveness across diverse domain groups. Overall, the PFL model demonstrates superior performance in the majority of domain groups, showcasing its potential for enhancing speaker identification tasks in various acoustic environments.

Evaluation of Continual Personalized Federated Learning

Table 5 presents the performance of Continual Personalized Federated Learning (C-PFL). We evaluate the model across all episodes and stages throughout the three training stages. It is worth noting that we use the same evaluation sets across all training stages, which serves as a good way to assess how different learning methods adapt to ever-changing new data and are influenced by the deletion of previous data from storage. Remarkably, our C-PFL method effectively integrates information from the data while enabling knowledge generalization and preventing catastrophic forgetting.

Upon examining the results, we observe a consistent improvement in the mean EER during the training of each stage, indicating enhanced performance. Furthermore, the standard deviation decreases, signifying a more uniform improvement across all evaluation domain groups. This trend underscores the robustness

Table 4 Evaluation performance of Centeralized training and PFL training on the speaker identification task.

Groups	Centeralized [14]				PFL (Ours)			
	ACC	Precision	F1-score	AUC	ACC	Precision	F1-score	AUC
G1	**0.9909**	**0.9912**	**0.9908**	**0.9996**	0.9664	0.9680	0.9654	0.9986
G2	0.6691	0.7945	0.6674	0.9785	**0.7909**	**0.8138**	**0.7845**	**0.9862**
G3	**0.9747**	**0.9759**	**0.9746**	**0.9987**	0.9379	0.9412	0.9368	0.9977
G4	0.3633	0.5805	0.3634	0.9254	**0.6802**	**0.7151**	**0.6764**	**0.9695**
G5	**0.9696**	**0.9717**	**0.9696**	**0.9982**	0.9421	0.9458	0.9416	0.9977
G6	0.4946	0.7006	0.5178	0.9439	**0.7337**	**0.7819**	**0.7289**	**0.9802**
G7	**0.5837**	**0.6658**	**0.5606**	0.9527	0.5565	0.6323	0.5359	**0.9564**
G8	0.2136	0.4647	0.2165	0.8485	**0.3424**	**0.4301**	**0.3139**	**0.9092**
G9	0.4904	0.6035	0.4809	0.9266	**0.5088**	**0.5672**	**0.4877**	**0.9460**
G10	0.2364	0.4085	0.2309	0.8540	**0.3821**	**0.4584**	**0.3685**	**0.9177**
G11	0.4630	0.6267	0.4800	0.9123	**0.5257**	**0.5823**	**0.5160**	**0.9469**
G12	0.2396	0.4812	0.2440	0.8448	**0.4101**	**0.4917**	**0.4055**	**0.9244**
Mean(μ)	0.5574	0.6887	0.5580	0.9319	**0.6481**	**0.6940**	**0.6384**	**0.9609**

Table 5 Evaluation performance of C-PFL in continual learning settings.

Evaluation Groups	Canonical model EER	C-PFL Stage1 EER	C-PFL Stage2 EER	C-PFL Stage3 EER
G1	2.85	**4.737**	5.418	5.313
G2	33.06	**13.844**	13.828	**13.844**
G3	4.81	**6.856**	7.316	7.388
G4	39.23	19.751	19.158	**18.523**
G5	5.51	**6.34**	6.779	6.773
G6	37.80	16.605	16.876	**15.695**
G7	16.25	15.925	15.529	**15.696**
G8	41.29	24.111	23.69	**22.913**
G9	20.61	17.287	**17.053**	**17.09**
G10	37.61	24.512	23.69	**22.913**
G11	24.45	17.484	**17.338**	17.52
G12	37.10	22.649	22.434	**22.117**
mean(μ)	25.0475	14.965	14.922	**14.678**
std(σ)	14.214	6.937	6.399	**6.152**

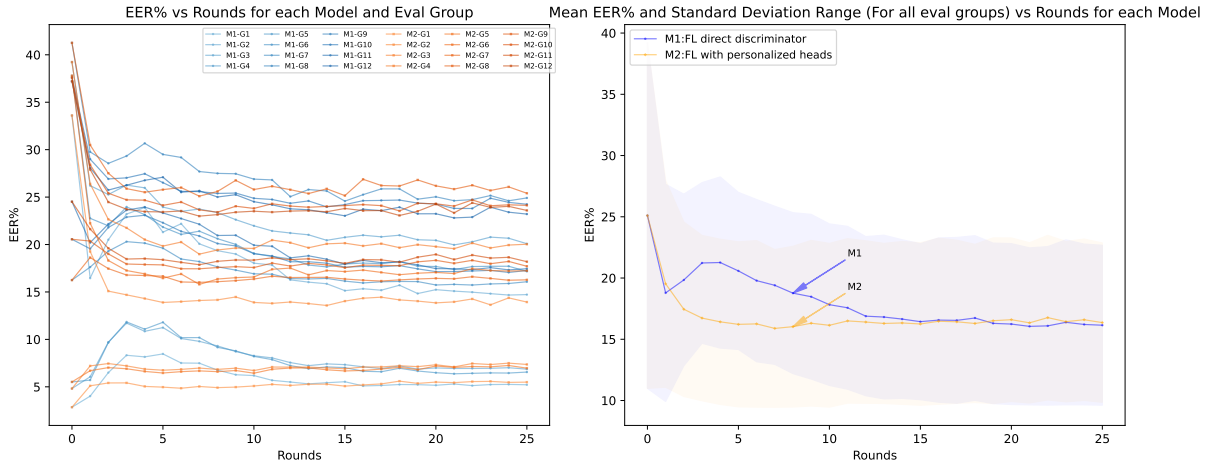

Figure 5 Comparing the error rate between the FL-classical and the PFL Type-A.

of the C-PFL method in adapting to various domain datasets and demonstrates its potential for real-world continual learning applications. The model effectively generalizes to new knowledge while avoiding catastrophic forgetting.

In Figure 6, we compare the performance of the standard C-PFL strategy with our enhanced C-PFL strategy across each stage of the learning process. The standard strategy appears to struggle with retaining previously acquired knowledge and generalizing it to future learning stages. In contrast, the enhanced strategy demonstrates superior performance in continual learning scenarios.

This improvement can be attributed to the design of the proposed C-PFL method, which effectively balances the retention of prior knowledge with the acquisition of new information by randomly shifting the weight designation in the output classifier of the PFL module. As a result, our enhanced strategy exhibits a more stable learning curve, ensuring consistently better performance across various stages and domain datasets. This finding highlights the advantages of employing the C-PFL with an enhanced strategy approach in real-world continual learning speaker recognition applications, where knowledge retention and generalization are critical for achieving reliable and robust performance.

In Figure 9, we compare the impact of different training data incoming sequences on the continual learning process to assess the influence that varying data sequences may have on the results. We examine two scenarios: one where we use Stage 2 datasets first, followed by Stage 3 datasets, and another where we reverse the order. Our findings demonstrate that the proposed strategy remains effective regardless of the data sequence used, as evidenced by the EER plots in the figure.

In both cases, the EER on evaluation sets consistently improves at each stage, demonstrating that the system effectively leverages the knowledge acquired from previous stages to enhance subsequent training, regardless of the data incoming sequence. This observation underscores the robustness of the proposed C-PFL method.

Complementary Experiments

We perform supplementary experiments to provide a deeper understanding and analysis of the characteristics of our proposed PFL method.

Figure 7 displays the loss tracking for various selected groups of our training data when implementing our proposed PFL Type-A system. Different groups necessitate varying numbers of training steps within a single combination episode, causing some clients to wait for others to finish their training.

At the beginning of each episode, the loss initially increases to a local peak but converges rapidly compared to the previous episode, resulting in a lower loss value. This behavior illustrates the successful knowledge sharing and transfer between clients at each episode, enabling effective learning and adaptation across clients from multiple heterogeneous domains. The moving average mean and the standard deviation range shown in the plot indicate that all groups achieve good convergence within several hundred steps. This consistent convergence across various groups highlights the effectiveness of our proposed PFL approach in handling diverse data and domain conditions, ultimately improving the overall performance in speaker recognition tasks.

Figure 8 presents the t-SNE visualization of the embedding space for both personalized federated training (PFL) and centralized training strategies. Each point represents an embedding from an utterance, with each cluster corresponding to a speaker class, while different colors indicate speakers from distinct domain client groups. It is evident that both PFL and centralized training methods encounter challenges in distinguishing some speakers and forming well-defined clusters for certain speaker IDs. This difficulty arises due to the severe interference caused by the diverse room acoustic conditions experienced by some groups.

Nevertheless, a significant difference between the two methods can be observed. Centralized training seems to blend domain information with speaker class information, resulting in a decline in performance compared to the PFL training. This confusion is particularly evident in the highlighted area of the t-SNE plot, where the speaker classes and domain groups are evidently entangled. In contrast, the PFL strategy exhibits better separation of speaker classes and domain information, leading to improved overall performance in speaker recognition tasks. This can be attributed to the personalized module constraining domain-specific information within a single client training through feature projection, thus avoiding interference with the global model training. Consequently, PFL inherently emerges as a suitable method to effectively address the challenges posed by complex room acoustics and diverse domain conditions, ultimately enhancing the robustness and generalizability of speaker recognition systems.

Figure 10 presents the results of a common problem in Federated Learning where, at each combination time point, not all clients participate in the aggregation process. We investigate the effects of varying combination ratios for the proposed PFL system, including 1.0 (normal case), 0.7, and 0.3. As the combination ratio decreases, we observe a decline in performance, yet all scenarios still outperform the centralized training baseline (represented by the grey dashed line). The minimum EER dotted lines for different systems are displayed in the plot to emphasize this trend.

These findings emphasize the importance of the combination ratio in achieving optimal performance with FL. Despite the performance degradation caused by a lower combination ratio, the proposed federated learning method remains effective, consistently outperforming the centralized training method. Although the performance is reasonably good compared to the baseline, the influence of partial combination, unsynchronized combination, and the waiting time for clients on PFL

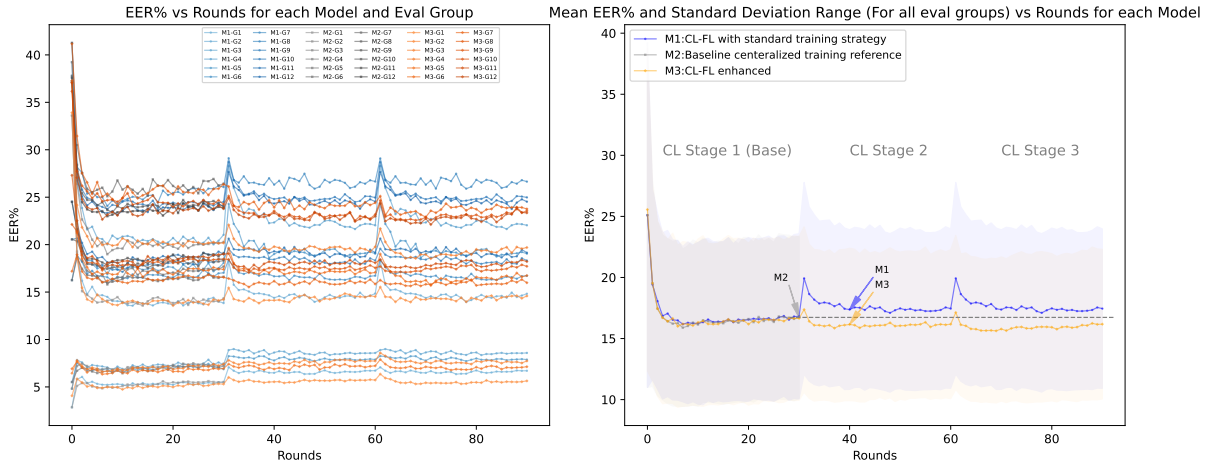

Figure 6 Comparison of error rates between C-PFL with and without enhanced training strategy, and the reference performance of single-stage non-continual PFL training.

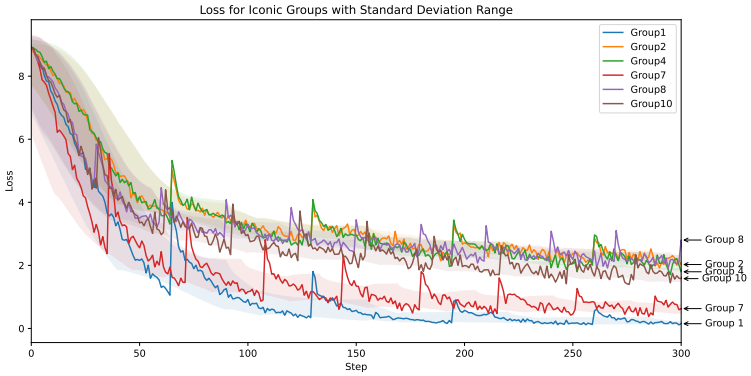

Figure 7 Tracking the loss learning curve for PFL training. Loss values for selected representative groups, along with their standard deviation range. The plot illustrates the loss values for Groups 1, 2, 4, 7, 8, and 10 during the first 300 steps of training.

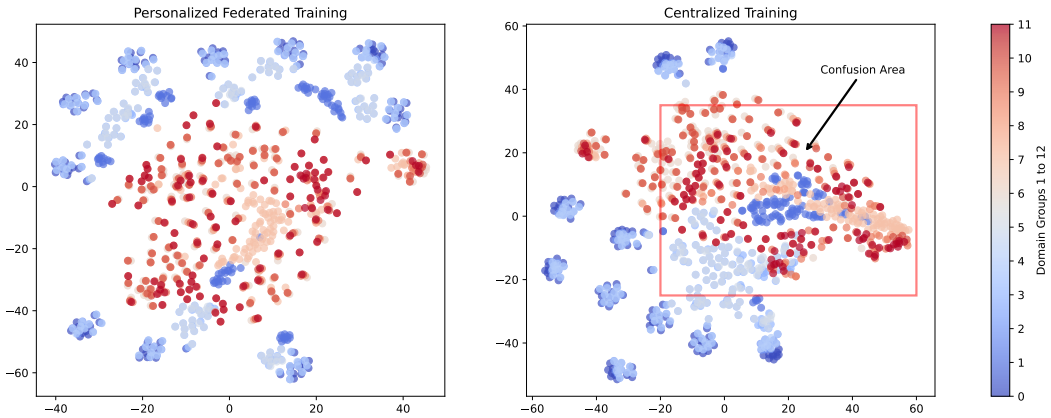

Figure 8 Comparing the embedding plot between the PFL and centeralized training.

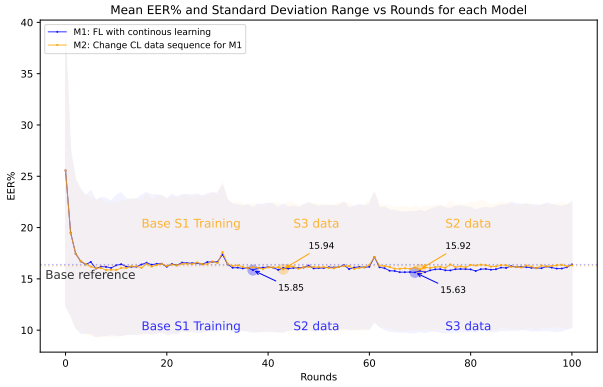

Figure 9 Comparison of error rates for C-PFL with different training data input sequences.

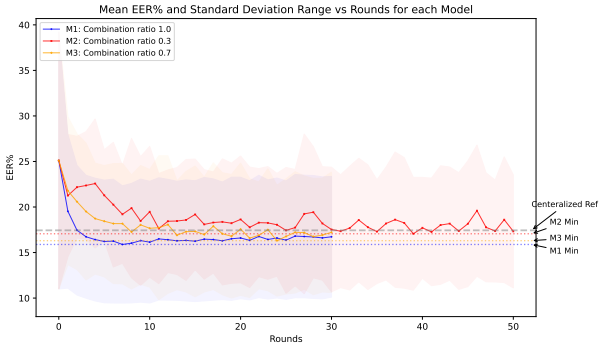

Figure 10 Comparison of error rates for PFL using various combination probabilities.

still remains an open problem, requiring further research.

Discussion and Future Considerations

In the present study, we have introduced a novel Personalized Federated Learning (PFL) system, engineered explicitly for speaker recognition tasks across various heterogeneous domains. The system has shown considerable advancements over existing baseline models. However, several aspects necessitate further exploration and discussion.

Future Research and Application

Moving forward, we aim to delve deeper into more advanced Federated Learning algorithms, such as those employing clustering-based FL methods [48]. Additionally, we intend to explore more pragmatic on-device learning algorithms, specifically those utilizing unsupervised streaming audio data. Our future research will also concentrate on crafting specific federated learning algorithms tailor-made for speaker diarization systems [12].

System Design and Reliability

The development of FL algorithms is intrinsically linked to system design; therefore, their integration into real-world systems demands careful consideration of numerous factors. Apart from the FL combination ratios discussed earlier, one significant element to consider is the implementation of asynchronous federated systems [49]. These systems cater to variations in client device timing prior to the aggregation in each training round, addressing potential downgrades in model accuracy due to disparities in device resources.

As we contemplate incorporating these features into tangible systems, the emphasis on system reliability intensifies. To mitigate reliability concerns, we can apply insights from Boudi et al.'s 2023 study [50], which showcases a robust and error-free career agent created using Deep Reinforcement Learning in tandem with formal verification. Moreover, the methodology presented by Ait-Ameur et al. (2023) [51] provides an excellent illustration of employing formal methods to handle complex cyber-physical systems. With the application of the Event-B formal specification language to model, verify, and refine these systems, we are well-equipped to design and verify the reliability of our FL-based speaker recognition system, accounting for variables such as device training latency and transmission latency, ensuring the system meets certain criteria.

Significance to AI, Speech and NLP Fields

Our research contributes to the wider fields of Artificial Intelligence (AI), speech, and Natural Language Processing (NLP). In these research domains, one prevailing trend involves the centralized training and pretraining of large-scale models for speech and NLP tasks, such as the Whisper model for speech recognition developed by OpenAI [52]. These models primarily aim to enhance generalization across diverse scenarios. Another emerging trend is the development of models that emphasize customization and personalization, as exemplified by Lin et al. (2023) [53], which cater to the unique complexities of Chinese language processing to improve NLP tasks.

We posit that by applying federated learning, starting from a well-pretrained model, we can concurrently enhance the model's generalization and personalization capabilities. This process is expedited and made seamless by the continual integration of wide-spread, privacy-sensitive data that would normally be unavailable or challenging to manage. Notably, in line with recent trends, the future integration of high-performance and explainable AI, as underscored by Bride et al. (2021) [54], presents an intriguing prospect for further enhancing our system's robustness and transparency. Consequently, this combined approach paves the way

for the steady evolution of our model, thereby advancing towards our aspiration of crafting a robust, explainable, and lifelong learning system in the future.

Conclusion

This study has demonstrated the effectiveness of Personalized Federated Learning in the field of speaker recognition. By facilitating the training of speaker recognition models using supervised speaker data stored on various heterogeneous domain silos, the proposed system has exhibited promising results in both speaker verification and speaker identification tasks. Moreover, the learning of the client-dependent personalized module has allowed for better adaptation to diverse scenarios.

Our simulations and evaluations, based on room acoustics software, have underscored the advantages of using PFL in heterogeneous domain adaptation scenarios. Comparisons between various Federated Learning methods, centralized training, and seperated training have revealed that PFL outperforms the other methods, a finding not reported in previous research.

Furthermore, PFL can be effectively integrated with continual learning settings, with the Continual Personalzied Ferderated Learning method showcasing strong performance as the training stages progress. By employing a random prototype casting training strategy, the C-PFL with enhanced strategy has proven advantageous.

In summary, the findings of this study endorse the adoption of Personalized Federated Learning as a valuable approach to speaker recognition tasks. This approach offers improvements in performance for heterogeneous domain adaptation compared to classical transfer learning methods for speaker recognition. Additionally, it presents a new way of learning in a continual, collaborative manner that brings valuable traits such as data security, training-costs saving, and adaptability to ever-changing data in real-world scenarios.

Declarations

Acknowledgements

We would like to express our gratitude to the various funding agencies that have supported this work.

Funding

This work was supported in part by the National High-Quality Program grant TC220H07D, the National Natural Science Foundation of China (NSFC) under Grant 61871262, 62071284, and 61901251, the National Key R&D Program of China grants 2022YFB2902000, Key-Area Research and Development Program of Guangdong Province grant 2020B0101130012, Foshan Science and Technology Innovation Team Project grant FS0AAKJ919-4402-0060.

Abbreviations

Not applicable

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the authors upon reasonable request. Any additional materials, such as code or supplementary information, can also be provided by the authors upon request. The relevant code and data that support the findings of this study are also available at https://github.com/bicbrv/FedSPK following the date of publication.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Zhiyong Chen (the first author) designed the study, performed the experiments, and contributed to the analysis and interpretation of the data. Prof. Shugong Xu (the corresponding author) contributed to the conception and overall design of the study, as well as supervised the research project. All authors participated in drafting and revising the manuscript, and all authors have read and approved the final version of the manuscript.

Authors' information

- Zhiyong Chen received his M.Eng. degree in Communication Engineering from Shanghai University, China, in 2021. He is currently pursuing a Ph.D. in Information and Communication Engineering at the same institution. His research interests include speaker recognition, speech recognition, acoustics, and emerging machine learning paradigms.
- Shugong Xu (Fellow, IEEE) received his master's degree in pattern recognition and intelligent control and a Ph.D. degree in EE from Huazhong University of Science and Technology, China. He is a Professor at Shanghai University, where he was the Founding Head of the Shanghai Institute for Advanced Communication and Data Science. Before joining Shanghai University, he held positions at Intel Labs, Huawei Technologies, Sharp Laboratories of America, and conducted research at the City College of New York, Michigan State University, and Tsinghua University.
  Prof. Xu has published over 160 peer-reviewed research papers, holds more than 60 U.S. and China patents, and has made significant contributions to international standards such as IEEE 802.11, 3GPP LTE, and DLNA. His research interests include 6G wireless communication systems, machine learning, pattern recognition, and AI-enabled embedded systems.
  In recognition of his work, Prof. Xu was awarded the 'National Innovation Leadership Talent' by the China Government in 2013 and elevated to IEEE Fellow in 2015. He also received the 2017 Award for Advances in Communication from the IEEE Communications Society.

Author details

[1]School of Communication and Information Engineering, Shanghai University, Shanghai, China. [2]School of Communication and Information Engineering, Shanghai University, Shanghai, China.

References

1. Bai, Z., Zhang, X.-L.: Speaker Recognition Based on Deep Learning: An Overview. Neural Networks **140**, 65–99 (2021)
2. Tu, Y., Lin, W., Mak, M.-W.: A Survey on Text-Dependent and Text-Independent Speaker Verification. IEEE Access **10**, 99038–99049 (2022)
3. OpenMined. https://www.openmined.org/
4. Woubie, A., Bäckström, T.: Federated Learning for Privacy Preserving On-Device Speaker Recognition. In: 2021 ISCA Symposium on Security and Privacy in Speech Communication, pp. 1–5 (2021)
5. Tan, A.Z., Yu, H., Cui, L., Yang, Q.: Towards Personalized Federated Learning. IEEE Transactions on Neural Networks and Learning Systems, 1–17 (2022)
6. Wang, L., Zhang, X., Su, H., Zhu, J.: A Comprehensive Survey of Continual Learning: Theory, Method and Application. ArXiv preprint arXiv:2302.00487 (2023)
7. Scheibler, R., Bezzam, E., Dokmanić, I.: Pyroomacoustics: A Python package for audio room simulations and array processing algorithms. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 351–355 (2018)

8. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing **19**(4), 788–798 (2011)
9. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-Vectors: Robust DNN Embeddings for Speaker Recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5329–5333 (2018)
10. Desplanques, B., Thienpondt, J., Demuynck, K.: ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In: Interspeech 2020, pp. 3830–3834 (2020)
11. Zhao, M., Ma, Y., Liu, M., Xu, M.: The SpeakIn System for VoxCeleb Speaker Recognition Challange 2021. ArXiv preprint arXiv:2109.01989 (2021)
12. Park, T.J., Kanda, N., Dimitriadis, D., Han, K.J., Watanabe, S., Narayanan, S.: A Review of Speaker Diarization: Recent Advances with Deep Learning. Computer Speech & Language **72**, 101317 (2022)
13. Wang, Z., Yao, K., Li, X., Fang, S.: Multi-Resolution Multi-Head Attention in Deep Speaker Embedding. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6464–6468 (2020)
14. Wang, R., Ao, J., Zhou, L., Liu, S., Wei, Z., Ko, T., Li, Q., Zhang, Y.: Multi-View Self-Attention Based Transformer for Speaker Recognition. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6732–6736 (2022)
15. Ding, S., Chen, T., Gong, X., Zha, W., Wang, Z.: AutoSpeech: Neural Architecture Search for Speaker Recognition. ArXiv preprint arXiv:2005.03215 (2020)
16. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., Wei, F.: WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. IEEE Journal of Selected Topics in Signal Processing **16**(6), 1505–1518 (2022)
17. Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021)
18. Lee, K.A., Wang, Q., Koshinaka, T.: The CORAL+ Algorithm for Unsupervised Domain Adaptation of PLDA. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5821–5825 (2019)
19. Li, L., Zhang, Y., Kang, J., Zheng, T.F., Wang, D.: Squeezing Value of Cross-Domain Labels: A Decoupled Scoring Approach for Speaker Verification. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5829–5833 (2021)
20. Wang, Q., Okabe, K., Lee, K.A., Koshinaka, T.: A Generalized Framework for Domain Adaptation of PLDA in Speaker Recognition. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6619–6623 (2020)
21. Iman, M., Rasheed, K., Arabnia, H.R.: A Review of Deep Transfer Learning and Recent Advancements. Technologies **11**(2), 40 (2023)
22. Bhattacharya, G., Monteiro, J., Alam, J., Kenny, P.: Generative Adversarial Speaker Embedding Networks for Domain Robust End-to-End Speaker Verification. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6226–6230 (2019)
23. Rohdin, J., Stafylakis, T., Silnova, A., Zeinali, H., Burget, L., Plchot, O.: Speaker verification using end-to-end adversarial language adaptation. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6006–6010 (2019)
24. Wang, Z., Hansen, J.H.L.: Multi-source Domain Adaptation for Text-independent Forensic Speaker Verification. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30**, 60–75 (2021)
25. Kang, J., Liu, R., Li, L., Cai, Y., Wang, D., Zheng, T.F.: Domain-Invariant Speaker Vector Projection by Model-Agnostic Meta-Learning. ArXiv preprint arXiv:2005.11900 (2020)
26. Sarfjoo, S., Madikeri, S., Motlicek, P., Marcel, S.: Supervised domain adaptation for text-independent speaker verification using limited data. In: Interspeech 2020, pp. 3815–3819 (2020)
27. Du, C., Han, B., Wang, S., Qian, Y., Yu, K.: SynAug: Synthesis-Based Data Augmentation for Text-Dependent Speaker Verification. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5844–5848 (2021)
28. Huang, H., Xiang, X., Zhao, F., Wang, S., Qian, Y.: Unit Selection Synthesis Based Data Augmentation for Fixed Phrase Speaker Verification. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5849–5853 (2021)
29. Taherian, H., Wang, Z.-Q., Chang, J., Wang, D.: Robust Speaker Recognition Based on Single-Channel and Multi-Channel Speech Enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 1293–1302 (2020)
30. Mošner, L., Matějka, P., Novotný, O., Černocký, J.H.: Dereverberation and Beamforming in Far-Field Speaker Recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5254–5258 (2018)
31. Zheng, N., Li, N., Wu, B., Yu, M., Yu, J., Weng, C., Su, D., Liu, X., Meng, H.: A Joint Training Framework of Multi-Look Separator and Speaker Embedding Extractor for Overlapped Speech. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6698–6702 (2021)
32. Ma, H., Yi, J., Tao, J., Bai, Y., Tian, Z., Wang, C.: Continual Learning for Fake Audio Detection. ArXiv preprint arXiv:2104.07286 (2021)
33. Sustek, M., Sadhu, S., Hermansky, H.: Dealing with Unknowns in Continual Learning for End-to-end Automatic Speech Recognition. In: Interspeech 2022, pp. 1046–1050 (2022)
34. Yang, M., Lane, I., Watanabe, S.: Online Continual Learning of End-to-End Speech Recognition Models. In: Interspeech 2022, pp. 2668–2672 (2022)
35. He, C., Shah, A.D., Tang, Z., Sivashunmugam, D.F.N., Bhogaraju, K., Shimpi, M., Shen, L., Chu, X., Soltanolkotabi, M., Avestimehr, S.: FedCV: A Federated Learning Framework for Diverse Computer Vision Tasks. ArXiv preprint arXiv:2111.11066 (2021)
36. Zhu, H., Wang, J., Cheng, G., Zhang, P., Yan, Y.: Decoupled Federated Learning for ASR with Non-IID Data. In: Interspeech 2022, pp. 2628–2632 (2022)
37. Gao, Y., Parcollet, T., Zaiem, S., Fernandez-Marques, J., de Gusmao, P.P.B., Beutel, D.J., Lane, N.D.: End-to-End Speech Recognition from Federated Acoustic Models. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7227–7231 (2022)
38. Jia, J., Mahadeokar, J., Zheng, W., Shangguan, Y., Kalinli, O., Seide, F.: Federated Domain Adaptation for ASR with Full Self-Supervision. In: Interspeech 2022, pp. 536–540 (2022)
39. Gao, Y., Fernandez-Marques, J., Parcollet, T., Mehrotra, A., Lane, N.: Federated Self-supervised Speech Representations: Are We There Yet? In: Interspeech 2022, pp. 3809–3813 (2022)
40. Tomashenko, N., Mdhaffar, S., Tommasi, M., Estève, Y., Bonastre, J.-F.: Privacy Attacks for Automatic Speech Recognition Acoustic Models in A Federated Learning Framework. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6972–6976 (2022)
41. Li, X.-C., Tang, J.-L., Song, S., Li, B., Li, Y., Shao, Y., Gan, L., Zhan, D.-C.: Avoid Overfitting User Specific Information in Federated Keyword Spotting. In: Interspeech 2022, pp. 3869–3873 (2022)
42. Hard, A., Partridge, K., Chen, N., Augenstein, S., Shah, A., Park, H.J., Park, A., Ng, S., Nguyen, J., Lopez-Moreno, I., Mathews, R., Beaufays, F.: Production federated keyword spotting via distillation, filtering, and joint federated-centralized training. In: Interspeech 2022, pp. 76–80 (2022)
43. Feng, T., Narayanan, S.: Semi-FedSER: Semi-supervised Learning for Speech Emotion Recognition On Federated Learning using Multiview Pseudo-Labeling. In: Interspeech 2022, pp. 5050–5054 (2022)
44. Granqvist, F., Seigel, M., van Dalen, R., Cahill, A., Shum, S., Paulik, M.: Improving on-device speaker verification using federated learning with privacy. ArXiv preprint arXiv:2008.02651 (2020)
45. Wang, Y., Song, Y., Jiang, D., Ding, Y., Wang, X., Liu, Y., Liao, Q.: FedSP: Federated Speaker Verification with Personal Privacy Preservation. In: Algorithms and Architectures for Parallel Processing, Cham, pp. 462–478 (2021)

46. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: Deep Speaker Recognition. In: Interspeech 2018, pp. 1086–1090 (2018)
47. Fan, Y., Kang, J., Li, L., Li, K., Chen, H., Cheng, S., Zhang, P., Zhou, Z., Cai, Y., Wang, D.: CN-CELEB: a challenging Chinese speaker recognition dataset. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7604–7608 (2020)
48. Duan, M., Liu, D., Ji, X., Liu, R., Liang, L., Chen, X., Tan, Y.: FedGroup: Efficient federated learning via decomposed similarity-based clustering. In: 2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), pp. 228–237 (2021)
49. Xu, C., Qu, Y., Xiang, Y., Gao, L.: Asynchronous federated learning on heterogeneous devices: A survey. ArXiv preprint arXiv:2109.04269 (2023)
50. Boudi, Z., Wakrime, A.A., Toub, M., Haloua, M.: A deep reinforcement learning framework with formal verification. Formal Aspects of Computing **35**(1), 1–17 (2023)
51. Aït-Ameur, Y., Bogomolov, S., Dupont, G., Iliasov, A., Romanovsky, A., Stankaitis, P.: A refinement-based formal development of cyber-physical railway signalling systems. Formal Aspects of Computing **35**(1), 1–1 (2023)
52. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning (PRML), pp. 28492–28518 (2023)
53. Lin, M., Xu, Y., Cai, C., Ke, D., Su, K.: A lattice-transformer-graph deep learning model for chinese named entity recognition. Journal of Intelligent Systems **32**(1), 20222014 (2023)
54. Bride, H., Cai, C.-H., Dong, J., Dong, J.S., Hóu, Z., Mirjalili, S., Sun, J.: Silas: A high-performance machine learning foundation for logical reasoning and verification. Expert Systems with Applications **176**, 114806 (2021)