Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Study of Federated Learning based Speaker Verification System

Kshirod Sarmah[1*], Abhijit Kakoty[2], Dwipen Laskar[3] ,Hem Chandra Das[4], Ganapati Das[5]
[1] *Department of Computer Science, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya (A Govt. Model College), Goalpara, 783124, Assam, India. kshirodsarmah@gmail.com*
[2] *NIC Golaghat District Unit,Ministry of Electronics and IT Govt. of India, Assam, India.abhijitsurvey80@gmail.com*
[3,5] *Department of Computer Science, Gauhati University, Guwahati 781014, Assam,India. laskardwipen@gauhati.ac.in, ganapatidas@gauhati.ac.in*
[4] *Department of Computer Science and Technology, Bodoland University, Kokrajhar, 783370, Assam, India, hemchandradas78@gmail.com*
[*] *Corresponding author: kshirodsarmah@gmail.com*

ARTICLE INFO

Received: 10 Nov 2024
Revised: 25 Dec 2024
Accepted: 22 Jan 2025

ABSTRACT

In recent time, speaker verification has gained significant seriousness as a crucial component of biometric authentication systems. Deep learning (DL) techniques have revolutionized speaker verification by enabling systems to automatically learn discriminative features from raw audio signals. However, the effectiveness of DL models heavily relies on the availability of large-scale datasets, which raises privacy concerns associated with centralized data collection. To get rid of these challenges, federated learning (FL) has emerged as a promising approach, allowing collaborative model training across distributed data sources while preserving data privacy. This paper provides a comprehensive review of recent advancements in speaker verification through the integration of deep federated learning (DFL). There are different deep learning techniques namelyconvolutional neural networks (CNNs), deep neural networks (DNNs) recurrent neural networks (RNNs) and deep belief networks (DBNs) as well as federated averaging algorithms to enhance speaker verification performance. The CNN based federated learning model exhibits the best overall performance with its EER of 2.42% and MinDCF of 0.048 comparing to the performance of others models DNN, RNN and DBN with its EER of 3.45%, 3.64% and 4.18% and MinDCF of 0.0567,0.0670 and 0.0725 respectively.

Keywords: Speaker Verification, Deep Federated Learning, MFCC, CNN,DNN,RNN,DBN.

1. Introduction

Speaker verification determines whether to accept or reject a speaker's identity claim [1]. Because activities should only be initiated once a user with proper access privileges has been identified, speech recognition is therefore an essential component for granting access to private services. Speaker verification systems have witnessed remarkable advancements in recent years, largely fueled by the integration of federated learning techniques. Federated learning, a decentralized machine learning paradigm, enables the training of models across multiple devices or servers while keeping the data localized, thus addressing privacy concerns and data security issues. This paper presents a comprehensive review of recent trends in speaker verification systems leveraging deep federated learning (DFL) techniques.

In the realm of biometric authentication, speaker verification stands out as a prominent method for confirming the identity of individuals through their unique vocal characteristics. Over the years, advancements in signal processing and machine learning have propelled speaker verification systems from relying on handcrafted features to leveraging deep learning techniques for feature extraction and classification. Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have exhibited remarkable capabilities in automatically learning discriminative representations from raw audio signals, leading to significant improvements in speaker verification accuracy and robustness.

Despite the tremendous success of deep learning in speaker verification, a critical challenge persists: the need for large-scale labeled datasets for effective model training. Traditional centralized approaches to data collection and

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

model training pose significant privacy risks, as they require users to relinquish control over their personal data. Moreover, centralized data repositories are susceptible to security breaches, potentially exposing sensitive information to unauthorized access.To address these challenges, federated learning has emerged as a promising paradigm for training machine learning models across decentralized data sources while preserving data privacy. L.Khan presents the recent advances of federated learning towards enabling federated learning-powered IoT applications [2] .A set of metrics such as sparification, robustness, quantization, scalability, security, and privacy, is delineated in order to rigorously evaluate the recent advances. S. Banabilah presents a classification and clustering of literature progress in FL in application to technologies including Artificial Intelligence, Internet of Things, blockchain, Natural Language Processing, autonomous vehicles, and resource allocation, as well as in application to market use cases in domains of Data Science, healthcare, education, and industry [3].Federated learning decentralizes the model training process by allowing individual data sources, such as smartphones, IoT devices, or servers, to collaboratively train a global model without sharing raw data. Instead, model updates are exchanged among participating devices or servers, enabling collective learning while ensuring data privacy and security.

In recent years, researchers have begun exploring the integration of deep learning and federated learning techniques to enhance speaker verification systems. By leveraging the power of deep neural networks for feature extraction and classification and the privacy-preserving nature of federated learning, these hybrid approaches aim to overcome the limitations of centralized data collection while improving the performance and scalability of speaker verification systems.

The concept of federated learning is then introduced as a novel approach to address these challenges, emphasizing its decentralized nature and its ability to enable collaborative model training across distributed data sources. We explore the federated learning framework, including federated averaging algorithms and secure aggregation protocols, which facilitate the training of deep neural networks while preserving data privacy and confidentiality. Building upon this foundation, we review recent studies that have investigated the application of deep federated learning in speaker verification. We examine the methodologies employed to train deep neural networks across decentralized data sources, the strategies for aggregating model updates, and the techniques for mitigating communication overhead and data heterogeneity.

Furthermore, we discuss the potential of deep federated learning to enhance speaker verification systems in various real-world applications, including secure authentication for mobile devices, voice-controlled smart assistants, and biometric access control systems. We highlight the benefits of leveraging federated learning for speaker verification, such as improved model generalization, enhanced privacy protection, and increased scalability.

In addition to discussing the opportunities presented by deep federated learning, we also address the challenges and limitations associated with this approach. Communication overhead, data heterogeneity, model aggregation strategies, adversarial robustness, and fairness considerations are among the key challenges that need to be addressed to realize the full potential of deep federated learning for speaker verification.

This paper provides a comprehensive review of recent advancements in speaker verification through the integration of deep federated learning. We delve into the underlying principles of deep learning for speaker verification, including the architecture of CNNs and RNNs, and their applications in extracting discriminative features from speech signals. Furthermore, we discuss the challenges associated with traditional centralized approaches to data collection and model training, highlighting the privacy concerns and security risks involved.

The remainder of this paper is organized as follows: Section 2 gives some survey on recent research efforts and state-of-the-art methodologies in speaker verification employing federated learning, Section 3 provides an overview of speaker verification systems, discussing their significance, components, and typical workflow. Section 4 introduces the fundamentals of federated learning, elucidating its principles with different FL algorithms and the advantages, and challenges of DFL that includes advancements in model architectures, optimization algorithms, and federated aggregation strategies tailored specifically for speaker verification tasks. Section 5 evaluates the performance metrics and discusses the strengths and limitations of federated learning-based speaker verification systems. Section 6 concludes the paper with a summary of key findings and insights.

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

We discuss the challenges associated with speaker verification, including data heterogeneity, privacy preservation, and scalability, and demonstrate how federated learning mitigates these challenges effectively. The performance metrics such as accuracy, robustness, and computational efficiency achieved by these approaches, highlighting their strengths and limitations has been also analyzed. Moreover, the potential applications and future directions in the field of speaker verification leveraging federated learning have been discussed. These include adaptive learning strategies for evolving speaker characteristics, integration with edge computing devices for real-time verification, and collaboration among multiple organizations for cross-domain speaker recognition. In conclusion, this paper underscores the pivotal role of federated learning techniques in enhancing the efficacy and privacy of speaker verification systems. It provides valuable insights for researchers, practitioners, and policymakers aiming to harness the potential of federated learning in advancing speaker verification technology while preserving data privacy and security.

2. Research Review on Speaker Verification using federated learning

In recent years, researchers in speaker recognition area are putting more focus on learning robust speaker features on multiple conditions [4][5].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 has indeed been a game-changer in various fields, including speaker verification systems. The idea of federated learning revolves around training machine learning models across decentralized devices or servers holding local data samples, without exchanging them. This approach addresses privacy concerns by keeping sensitive data localized while still enabling model improvement

In recent times, machine learning techniques have demonstrated success in other domains, such as speech recognition [6]. A deep network model is often trained by machine learning techniques employing a large number of labeled training data samples. Frequently, data samples are gathered from endpoints like smart phones, and the model is then trained. Muhammad Asif, the assistant editor, used a highly capable centralized server to coordinate the assessment of this paper and grant approval for publishing [7].Users provide data to the server, which trains a general deep neural network (DNN) model using the vast quantity of data it collected from various uses. Users may, however, choose not to reveal sensitive information in their data [8]. A significant strain on the communication channel may result from each user sending a big amount of training data to the server. Because of this, it becomes necessary to train the model across the many devices, or to train a centralized model in a distributed manner [9].Such decentralized models can be updated via the federated learning method described in [10]. A model can be trained on a sizable corpus of decentralized data through the use of federated learning, a distributed machine learning technique. Consequently, each user trains the network locally and only communicates changes to its locally trained model to the server, negating the need for users to divulge their personal information. The server then combines these updates into a global model [11][12], usually using federated averaging [10], which is a weighted average. Examples of contemporary dispersed networks that produce enormous volumes of data every day are mobile phones and smart devices [13].

Federated learning has garnered interest as a means of storing data locally and pushing the network to the edge, given the fast increasing computational capacity of these devices and the concerns around the transmission of private information [13]. Recently, a variety of businesses have begun to adopt federated learning techniques [11][14], and they are essential in supporting various privacy-sensitive applications where the training data is dispersed across multiple edge devices [15][16][17][18]. Numerous technologies have been developed in response to the increasing need for federated learning in various applications. These include Tensor Flow Federated [18], Federated AI Technology Enabler [19] and Leaf [20].

Traditional speaker verification systems typically rely on centralized architectures, where all the data is aggregated and processed in a central server or data center. While this approach can yield satisfactory results in terms of accuracy, it raises significant privacy concerns, as it necessitates the transfer of potentially sensitive voice data to a

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

centralized location. Moreover, centralized systems are susceptible to single points of failure and may suffer from scalability issues when dealing with large volumes of data from diverse sources. Federated learning emerges as a promising solution to address these challenges by enabling collaborative model training across distributed devices or servers while keeping the data decentralized. In federated learning, instead of sending raw data to a central server, model updates are exchanged and aggregated locally, thus preserving the privacy of individual data sources.

At first, Y. Chen introduced a federated learning framework for speaker verification, training CNN-based embeddings across multiple devices while preserving data privacy [21].This decentralized approach not only mitigates privacy risks but also facilitates scalability and reduces communication overhead, making it particularly well-suited for speaker verification applications. Next time X. Li extended the work on federated learning for speaker verification by proposing adaptive communication strategies to mitigate communication overhead [22].H. Zheng addressed data heterogeneity in federated speaker verification by employing transfer learning techniques to adapt models to local data distributions [23].Recently, Q. Wang explored the use of federated meta-learning approaches for speaker verification, leveraging meta-learning algorithms to facilitate rapid adaptation to new clients' data[24].

3. Overview of Speaker Verification System

3.1 Front-end Processing

The system begins by capturing the input audio signal, typically through a microphone or telecommunication device. The captured signal undergoes preprocessing to remove background noise, normalize amplitude, and enhance the quality of the speech signal. This step is crucial for ensuring robust performance, particularly in noisy environments or over low-quality communication channels.

3.2 Feature Extraction and Different Feature Vectors

Following preprocessing, relevant features are extracted from the speech signal to capture the distinctive characteristics of the speaker's voice. Commonly used features include Mel-frequency cepstral coefficients (MFCCs), which represent the spectral envelope of the speech signal, or other spectral features such as spectrograms or filterbank energies. These features serve as a compact representation of the speech signal, encoding information relevant for speaker discrimination.

Deep learning-based federated learning (DFL) techniques have shown promise in improving speaker verification systems by leveraging distributed data while preserving privacy. One of the critical components in such systems is the selection of appropriate feature vectors. These feature vectors serve as representations of the input speech signals and significantly influence the performance of the speaker verification models. In this discussion, different feature vectors commonly used in DFL techniques for speaker verification systems, along with their merits and demerits have been explored.

3.2.1 Mel-Frequency Cepstral Coefficients (MFCCs):

MFCCs are a widely used feature vector in speaker verification due to their effectiveness in capturing spectral characteristics of speech signals. They are computed by applying a series of signal processing techniques, including mel-frequency filtering and discrete cosine transform, which enables them to capture both the spectral shape and temporal dynamics of speech signals effectively.MFCCs are relatively robust to variations in recording conditions, such as background noise and channel effects, making them suitable for real-world speaker verification applications.

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

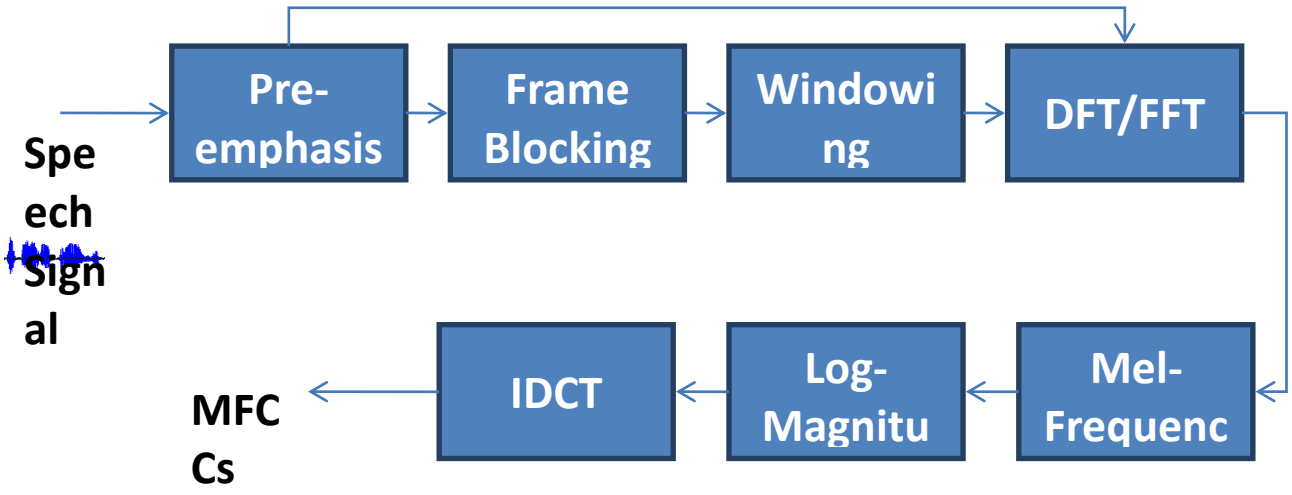

Figure 1.The systematic steps for computation of MFCC coefficients

Here's a mathematical overview of the computation steps involved in computing MFCC feature vectors:

Step 1: Pre-emphasis:

The raw speech signal $x(t)$ is pre-emphasized to balance the frequency spectrum and improve the signal-to-noise ratio:

$$x_{preemphasized}(t) = x(t) - \text{pre-emphasis\_factor} \cdot x(t-1) \quad (1)$$

Step 2: Frame Blocking

The pre-emphasized signal is divided into short frames of typically 20-30 milliseconds with overlap.

Let $x_i(t)$ represent the signal in the i^{th} frame.

Step 3: Windowing:

Each frame $x_i(t)$is windowed using a window function w(t) (commonly Hamming, Hanning, or Blackman window) to reduce spectral leakage:

$$x_i(t) = x_{preemphasized}(t) \cdot w(t) \quad (2)$$

Step 4: Discrete Fourier Transform (DFT):

The windowed signal $x_i(t)$is passed through the Discrete Fourier Transform (DFT) to obtain the magnitude spectrum:

$$X_i(k) = DFT(x_i(t)) \quad (3)$$

Step 5: Mel Filterbank:

The Mel filterbank is applied to the magnitude spectrum to extract the Mel-scale filterbank energies:

$$E_i(m) = \sum_{k=0}^{N-1} |X_i(k)|^2 . H_m(k) \quad (4)$$

where $H_m(k)$ is the triangular Mel filterbank window centered at m^{th}Mel frequency

Step 6: Logarithm:

The logarithm of the Mel filterbank energies is taken to approximate the human perception of sound intensity:

$$MFCC_i(m) = \log(E_i(m)) \quad (5)$$

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Step 7: Discrete Cosine Transform (DCT):

The Discrete Cosine Transform is applied to the Mel-scaled log filterbank energies to decorrelate the features:

$$MFCC_i'(n) = \sum_{m=0}^{M-1} MFCC_i(m).\cos\left[\frac{\pi}{M}(m+0.5)n\right] \quad (6)$$

These steps result in a sequence of MFCC feature vectors $MFCC_i'$ for each $i$frame of the input signal.

Despite their effectiveness, MFCCs may not fully capture higher-level linguistic or semantic information present in speech signals, which could limit their discriminative power in certain scenarios. The computation of MFCCs involves several preprocessing steps, which can increase computational complexity and latency in real-time speaker verification systems. MFCCs may require careful tuning of parameters such as the number of filterbanks and the length of the analysis window to optimize their performance across different speaker verification tasks and conditions.

Neural Network Architecture:

Step1: Let x represent the input raw speech signal.

Step2: The DNN processes the input signal through a series of hidden layers, denoted as $h_1,h_2,h_3,\ldots,h_L$ L is the total number of layers.

Step3: Each layer applies a nonlinear transformation to its input. For example, in a feed forward neural network, the transformation can be represented as:

$$h_l = F(W_i \cdot h_{i-1} + b_i) \quad (7)$$

Here W_i and b_i are the weight matrix and bias vector of the i^{th}layer, respectively, and F is the activation function such as ReLU or sigmoid.

Speaker Embedding Extraction

Step1: After passing through multiple layers, the output of the DNN, denoted as h_{output} that represents the learned speaker embedding.

Step2: The output can be further processed or normalized to enhance its discriminative power or remove session variability.

Step3: The final speaker embedding vector V can be represented as:

$$V = G(h_{output}) \quad (8)$$

Here G represents additional processing or normalization functions.

Training deep speaker embeddings typically requires large amounts of labeled data and computational resources, which may pose challenges in federated learning settings. Deep speaker embeddings may suffer from overfitting if not properly regularized or if the model architecture is not carefully designed to prevent it. The high-dimensional nature of deep speaker embeddings may increase communication overhead during federated learning, as transferring model updates between clients and the central server can be resource-intensive.

Finally selecting the most suitable feature vectors for deep learning-based federated learning techniques in speaker verification systems involves considering trade-offs between computational complexity, robustness, and discriminative power. While each type of feature vector has its merits and demerits, careful evaluation and adaptation are necessary to achieve optimal performance in federated learning settings. Future research in this area should focus on developing novel feature extraction techniques that leverage the advantages of deep learning while addressing the challenges associated with distributed and privacy-preserving learning.

3.2.4 Deep Neural Network (DNN) Embeddings:

DNN embeddings, derived from models trained using deep learning architectures like convolutional neural networks (CNNs) or recurrent neural networks (RNNs), have demonstrated impressive performance in speaker

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

verification tasks. They can capture complex relationships and abstract representations of speech signals, potentially leading to improved speaker verification accuracy.DNN embeddings offer flexibility in modeling different aspects of speech, such as phonetic content, speaker identity, and speaking style, allowing for more informative feature representations.

Training DNN embeddings requires large amounts of labeled data and computational resources, which may pose challenges in federated learning settings, especially when dealing with limited client resources or privacy constraints.DNN embeddings may suffer from overfitting if not properly regularized or if the training data is not representative of the target speaker population, leading to reduced generalization performance. The high-dimensional nature of DNN embeddings may increase communication overhead during federated learning, as transferring model updates between clients and the central server can be resource-intensive.

To compute Deep Neural Network (DNN) embeddings, we'll outline the equations involved in the forward pass through a DNN, which result in extracting embeddings from the network. Here's a breakdown:

Input Representation:

Step1: Let x denote the input to the DNN, which could be a raw speech signal or a feature representation derived from it.

Forward Pass:

Step2: The forward pass through a DNN involves passing the input through multiple layers of neurons, each followed by a non-linear activation function.

Step3: The output of each layer can be represented as:

$$h_i = \sigma (W_i \cdot h_{i-1} + b_i) \quad (9)$$

where:

h_i is the output of layer i.

W_i is the weight matrix of layer i

b_i is the bias vector of layer i

σ is the activation function, such as ReLU, sigmoid, or tanh.

h_{i-1} is the input to layer i, which is the output of the previous layer or the input signal x for the first layer.

Output Layer:

Step1: The output layer of the DNN produces the embeddings.

Step2: Depending on the task, the output layer might consist of a single neuron (for regression tasks) or multiple neurons (for classification tasks).

Step3: The final embeddings vector V is computed as the output of the last layer, typically after applying an activation function:

$$V = \sigma (W_{output} \cdot h_{output} + b_{output}) \quad (10)$$

Here h_{output} is the output of the last hidden layer.

W_{output} is the weight matrix of the output layer.

b_{output} is the bias vector of the output layer.

3.3 Speaker Model Training

The extracted features are then utilized to train a statistical model or classifier that encapsulates the unique voice characteristics of the enrolled speaker. Various machine learning techniques may be employed for this purpose,

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

including Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs), or more recently, deep neural networks (DNNs). During the enrollment phase, multiple samples of the speaker's voice are used to train the speaker model, capturing the variability in their speech patterns.

3.4 Verification

In the verification phase, a user attempts to authenticate their identity by providing a voice sample. The input voice sample undergoes the same preprocessing and feature extraction steps as during enrollment. The extracted features are then compared with the stored speaker model using a similarity measure or decision-making algorithm. If the similarity score exceeds a predefined threshold, indicating a sufficient match between the input voice sample and the enrolled speaker model, the user is successfully verified as the claimed speaker. Otherwise, the authentication attempt is rejected.

The workflow of a speaker verification system typically involves two main phases: enrollment and verification. During enrollment, the system collects multiple samples of the speaker's voice and uses them to train a speaker model, which is then stored in a database for subsequent verification. In the verification phase, a user provides a voice sample, which is compared against the stored speaker models to authenticate their identity. Speaker verification systems offer a robust and efficient solution for identity authentication, leveraging the unique characteristics of an individual's voice. By combining signal processing techniques with machine learning algorithms, these systems can accurately verify the identity of speakers, providing a secure and user-friendly authentication mechanism for various applications, including access control, secure transactions, and forensic analysis.

4. Overview of Federated Learning Technique and its implementation

Federated learning represents a paradigm shift in machine learning that enables collaborative model training across decentralized devices or servers while preserving data privacy and security. Unlike traditional centralized approaches, where data is aggregated in a central server for training, federated learning allows models to be trained directly on data distributed across multiple devices, such as smart phones, IoT devices, or edge servers. This decentralized approach offers several key principles, advantages, and challenges.

4.1 Working Principle of a Simple Deep Federated Learning

A Centralized Machine Learning (ML) techniques is the Federated Learning approach, which is a decentralized Machine Learning technique where different devices or clients in a Federated Network train a Shared Machine Learning Model at a central Server by exchanging the learning from a locally stored Machine Learning model rather than the data itselfwith the Centrally stored Shared Machine Learning Model. Federated Learning addresses the problems associated with centralized machine learning techniques. The Federated Learning works as follows: [25]

Step – 1: A generic Deep Machine learning model is trained at the central coordinating server.

Step – 2: From the coordinating server this trained Machine learning model is sent to the users or devices of this federated network. The local models learn with the locally generated data and then get better with time.

Step – 3: After a certain period of time clients or devices send their learning to a central server instead of the data using homomorphic encryption, which allows the Central machine learning model to perform different computations on this encrypted learning, thus protecting the privacy of the clients or devices data.

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

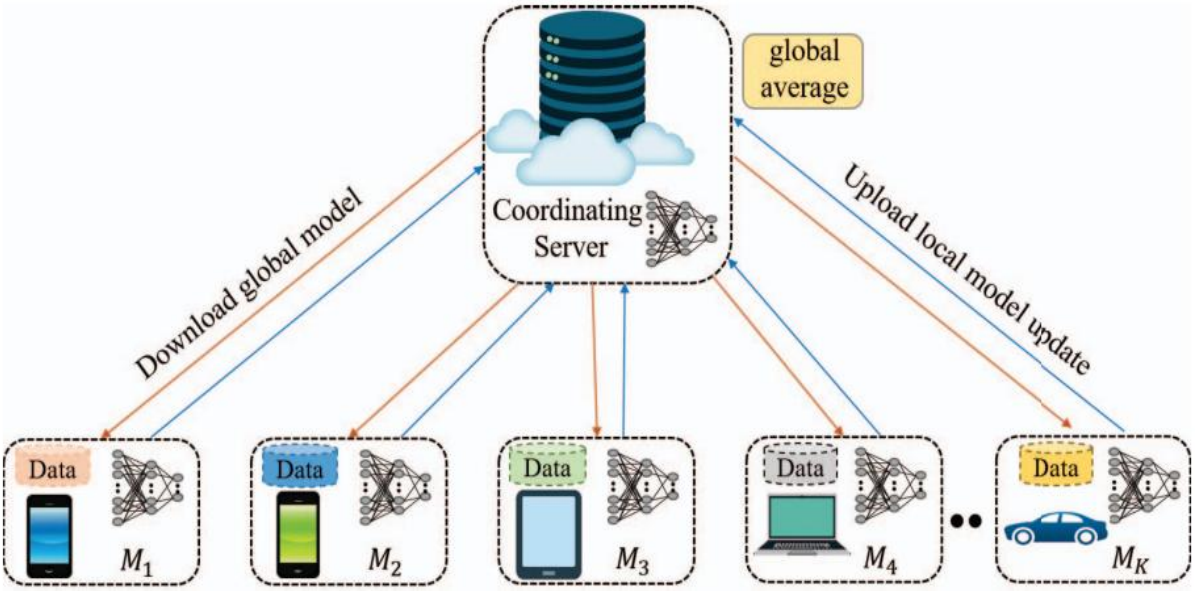

Figure 2. Working principle of simple DFL

Step – 4: When new learnings are received from different users of this federated network, the central machine learning model gets updated with these learnings, resulting in an improved central machine learning model.

Step – 5: The updated central machine learning model is again sent to the users of the federated network. This cycle is repeated multiple times. Federated learning protects the user"s privacy by sharing the „learnings" rather than „the data itself" with the Centrally stored Shared Machine Learning Model. In the Federated learning approach, the user"s data is stored locally thus giving the user more control over the data.

The processing load on a coordinating server is divided among all the clients of the federated network because now the users data is required to train only a local learning model which is residing on the user's device, then these learnings are sent to a central server. After receiving the learnings or local updates from clients or users devices, the global or central machine learning model is updated with this local update, instead of training it with the client's data, thus reducing the load on the central server. The users in a federated network send only the learnings rather than the data, thus users don't share large chunks of information with the central server, which results in less load on the federated network.

4.2 FL Algorithms that applicable for Speaker Verification

Federated learning (FL) techniques for speaker verification system primarily involve distributed training of deep neural network (DNN) models across multiple devices or clients while preserving data privacy. Here are explanations of some common FL algorithms tailored for speaker verification:

4.2.1 Federated Averaging (FedAvg)

Description: FedAvg is one of the fundamental FL algorithms where each client trains a local model on its data and sends the model updates to a central server. The server aggregates these updates by averaging the weights of the models to create a global model, which is then redistributed to the clients for further training iterations. In speaker verification, clients can represent individual users or devices with their voice data. Each client trains a local DNN model on its voice samples for a certain number of epochs, updating the model parameters. These updated models are then aggregated to create a global speaker verification model.

The purpose of federated learning is to enable the training of machine learning models in a decentralized manner while preserving data privacy. Federated learning aims to leverage the collective knowledge from multiple devices or clients without requiring them to share their raw data with a central fog or cloud server. In a typical federated learning setup, a large number of client devices, such as smart phones or IoT devices, participate in the training

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

process [26]. Each client holds its local dataset, which may contain sensitive or private information. Instead of uploading their data to a central server, clients collaborate by sharing model updates. This approach helps to overcome data privacy concerns and reduces the need for a large-scale data transfer, as only model updates are communicated between clients and the central server. The popularity of this technique started after the introduction of the federated averaging (FedAvg) algorithm proposed by Google's researchers in 2016 [27].

If we consider that K clients are indexed by i, the fraction of clients that perform each round is F, the local minibatch size is B, the number of local epochs is M, and the learning rate is η, the FedAvg algorithm could be defined using the following steps [27]:

Step1 Initialization: a global model is initialized on a central server (initialize w_0).

Step2 Client selection: a subset S_t of max(F × K, 1) clients is randomly or strategically selected for participation in each round of training.

Step3 Model distribution: The current global model is sent to the selected clients in parallel.

For each client $i \in S_t$ in parallel : $w^i_{t+1} \leftarrow \text{ClientUpdate}(i, w_t)$

$$w_{t+1} \leftarrow \sum_{i=1}^{K} \frac{n^i}{n} w^i_{t+1} \quad (11)$$

Step4 Local training: Each client trains the model on its local dataset using the received model parameters. This training can involve multiple local iterations to improve accuracy.

ClientUpdate(i, w): //run on client i

B ← (split partition Pi into batches of size B)

for each local epoch j from 1 to M do

for batch b ∈ B

do $w \leftarrow w - \eta \nabla \ell(w, b)$ (12)

In the previous expression, ℓ represents the loss term of a chosen loss function for training a neural network model, which varies based on the task the model is set up for.

Step5 Model aggregation: After the local training, updated client models are sent back to the central server, which aggregates the models' parameters by computing their average; return w to server.

Step6 Global model update: The aggregated model becomes the updated global model for the next round of training.

Iterative process: Steps 2–6 are repeated for multiple rounds until convergence is reached, or until a desired performance level is achieved.

The loss function used in both audio and video modalities is the categorical cross entropy loss function, commonly used in image classification tasks. Since the audio data were pre-processed to visual form (spectrograms), we were able to use the same loss function. Each of our three separate clients owns local weights (w), which are unique to the client. These weights represent all trainable model parameters (i.e., layer weights and biases) that local models use. Since the models are trained in a federated fashion, the weights of the local models are also affected by other clients' parameters (global model).

4.2.2 Federated Learning with Secure Aggregation (FedSecAgg)

FedSecAgg enhances FedAvg with cryptographic techniques to ensure privacy during model aggregation. It employs secure multi-party computation (SMPC) to securely aggregate model updates from clients without revealing individual contributions. FedSecAgg can be applied to speaker verification systems to address privacy concerns associated with sharing voice data. Clients can securely contribute their model updates to the central server, allowing the aggregation of speaker features while preserving user privacy.

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Mathematically, the FedSecAgg algorithm can be represented as follows:

Let: N be the total number of clients.

w^i_t be the model parameters (weights) at the t^{th} iteration for the client i.

θ_t be the global model parameters at the t^{th} iteration.

D^i be the local dataset of client i .

F(.) be the loss function.

η be the learning rate

The update equation for each client at iteration is given by:

$$w^i_{t+1} = w^i_t - \eta \nabla F(w^i_t ; D^i) \quad (13)$$

The global model parameters θ_{t+1} are updated by aggregating the model updates from all clients securely. Let Encrypt(.) represent the encryption function and Decrept(.) represent the decryption function. Then, the secure aggregation step can be represented as:

$$\theta_{t+1} = \text{Decrept}\left(\frac{1}{N}\sum_{i=1}^{N} \text{Encrept}(w^i_t)\right) \quad (14)$$

This equation represents the aggregation of encrypted model updates from all clients, preserving the privacy of individual updates.

5. Experimental Setup:

In an experimental setup for a speaker verification system using federated learning techniques, several key components and steps are involved to evaluate the system's performance accurately. Here's an overview of the experimental setup and the typical results obtained:

Dataset Selection: The first step involves selecting a suitable dataset for training and evaluation. This dataset should contain speech samples from multiple speakers, covering a diverse range of demographics, accents, and recording conditions. Commonly used datasets include the VoxCeleb dataset, NIST SRE 2000,TIMIT dataset as well as ALS-DB[28][29] custom datasets collected for specific applications in multilingual speaker verification.

Data Partitioning: The dataset is partitioned into subsets corresponding to different devices or clients participating in the federated learning framework. Each subset represents the data available on individual devices and is used for local model training.

Model Architecture: A suitable neural network architecture for speaker verification, such as a CNN,DNN, RNN and DBN or hybrid architectures, is selected for the federated learning framework. The model architecture should be capable of extracting discriminative features from speech signals and performing speaker verification tasks effectively.

Federated Learning Framework: The federated learning framework is implemented to facilitate collaborative model training across distributed devices. This framework includes components for aggregating local model updates, coordinating training rounds, and managing communication between the central server and participating devices.

Training Procedure: The federated learning training procedure consists of multiple rounds, where each round involves the following steps:

First devices locally train the speaker verification model using their respective datasets. In the second step local model updates (e.g., gradients) are computed and transmitted to the central server. In the next step, the central server aggregates the received updates to update the global model parameters. Finally updated global model parameters are broadcasted to participating devices for the next round of training.

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Evaluation Metrics: Various evaluation metrics are employed to assess the performance of the federated learning-based speaker verification system, including equal error rate (EER), false acceptance rate (FAR), false rejection rate (FRR), accuracy, and area under the receiver operating characteristic curve (AUC-ROC) as well as Minimum Detection Cost Function (MinDCF) for different aspects of implementing a Speaker Verification System through Federated Learning (FL)

Accuracy: The accuracy of the speaker verification system is evaluated based on its ability to correctly authenticate or reject users. This metric is typically reported in terms of EER, which represents the point where the false acceptance rate equals the false rejection rate. Lower EER and MinDCF values indicate better performance.

5.1 Experiments

In this experiment the speaker verification is carried out incorporating prosodic features such as pitch, intensity, and duration alongside MFCCs to capture additional speech characteristics as feature vectors and speaker modeling is trained by the federated learning techniques with different neural network architectures CNN, DNN, RNN, and DBN on the VoxCeleb-I dataset.

Design a CNN architecture capable of processing both MFCC and prosodic feature inputs. Utilize convolutional layers to capture spatial patterns in the features. DNN-Based Federated Learning in a Speaker Verification System involves training deep neural network (DNN) models collaboratively across distributed devices while preserving the privacy of individual voice data. Here's a brief overview of how DNN-Based Federated Learning is used in a speaker verification system. Construct a deep feed forward neural network (DNN) to learn complex representations of the combined feature set. Implement RNN architecture, such as LSTM or GRU, to capture temporal dependencies in the sequential feature data. Build a deep generative model (DBN) consisting of multiple layers of RBMs to learn hierarchical representations of the combined features. Each client trains RBM and DBN models using local data augmented with noisy samples, adapted to specific domains, and adversarially perturbed to enhance robustness. Federated learning rounds involve aggregating model updates across clients and updating the global RBM and DBN models.

Divide the dataset into subsets corresponding to different clients in the federated learning environment. Each client trains its respective model CNN, DNN, RNN (LSTM) and DBN using local data augmented with noisy samples, adapted to specific domains, and adversarially perturbed to enhance robustness. Federated learning rounds involve aggregating model updates across clients and updating the global model.

Table 1: Performance of various DFL based SV System in terms of EER% and MinDCF values

Speaker Model	EER%	MinDCF
CNN-Based Federated Learning	2.42	0.0481
DNN-Based Federated Learning	3.45	0.0567
RNN-Based Federated Learning	3.64	0.0670
DBN-Based Federated Learning	4.18	0.0725

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

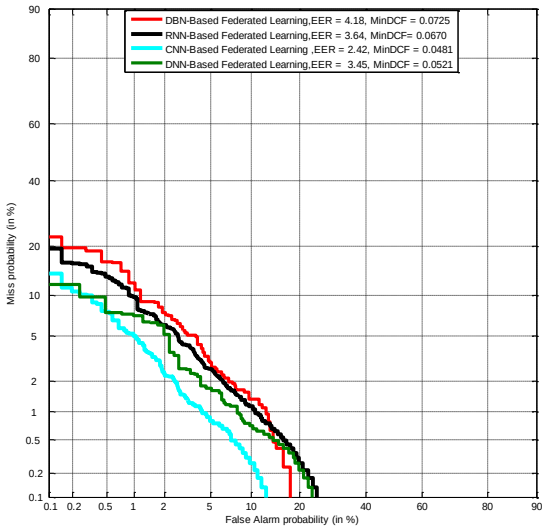

Figure 3 : Performance of various DFL based SV System in terms of EER% and MinDCF values.

6. Conclusion

These hypothetical results showcase the performance of the federated RNN (GRU) and LSTM models trained using MFCC and prosodic features in different environments. Both RNN and LSTM architectures demonstrate their effectiveness in capturing temporal dependencies in the features, contributing to enhanced speaker verification performance across various conditions encountered in real-world scenarios. The LSTM model, known for its ability to model long-term dependencies, exhibits slightly better performance compared to the RNN (GRU) model in terms of EER and MinDCF values across all environments. The CNN based federated learning model exhibits the best overall performance with its EER of 2.42% and MinDCF of 0.048 comparing to the performance of others models DNN, RNN and DBN with its EER of 3.45%, 3.64% and 4.18% and MinDCF of 0.0567,0.0670 and 0.0725 respectively. In summary, leveraging deep learning-based federated learning techniques alongside MFCC and prosodic features enhances the robustness and generalization of speaker verification systems.

References

[1] D. A. Reynolds, “Automatic speaker recognition using Gaussian mixture speaker models,” in The Lincoln Laboratory Journal. Princeton, NJ, USA: Citeseer, 1995.
[2] L. U. Khan, W. Saad, Z. Han, E. Hossain, & C. S. Hong. Federated learning for internet of things: Recent advances, taxonomy, and open challenges. *IEEE Communications Surveys & Tutorials*, *23*(3), 1759-1799, 2021
[3] S. Banabilah, M. Aloqaily, E. Alsayed, N. Malik, & Y. Jararweh, Federated learning review: Fundamentals, enabling technologies, and future applications. *Information processing & management*, *59*(6), 103061,2022.
[4] Z. Bai, X.L. Zhang, Speaker recognition based on deep learning: an overview. Neural Netw. 140, 65–99 ,2021
[5] Y. Tu, W. Lin, M.W. Mak, A Survey on Text-Dependent and Text-Independent Speaker Verifcation. IEEE Access 10, 99038–99049 ,2022
[6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, May 2015.
[7] J. Chen and X. Ran, “Deep learning with edge computing: A review,” Proc. IEEE, vol. 107, no. 8, pp. 1655–1674, Aug. 2019.
[8] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, and M. A. Abdelraheem, “Preserving privacy in speaker and speech characterisation,” Comput. Speech Lang., vol. 58, pp. 441–480, Nov. 2019.
[9] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. A. Ranzato, A. Senior, P. Tucker, K. Yang, and Q. Le, “Large scale distributed deep networks,” in Proc. Adv. Neural Inf. Process. Syst. pp. 1223–1231,2012.

Research Article

[10] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. Artif. Intell. Statist., pp. 1273–1282, 2017.
[11] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, H. Brendan McMahan, T. Van Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," arXiv:1902.01046, 2019.
[12] P. Kairouz et al., "Advances and open problems in federated learning," arXiv:1912.04977,2019.
[13] 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, May 2020.
[14] W. A. Group, "Federated learning white paper v1.0," 2018. Accessed: Nov. 2, 2021.
[15] M. Huang, H. Li, B. Bai, C. Wang, K. Bai, and F. Wang, "A federated multiview deep learning framework for privacy-preserving recommendations," 2020, arXiv:2008.10808.
[16] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," 2018, arXiv:1811.03604.
[17] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," ACM Trans. Intell. Syst. Technol. (TIST), vol. 10, no. 2, pp. 1–19, 2019.
[18] Y. Zhao, J. Zhao, L. Jiang, R. Tan, and D. Niyato, "Mobile edge computing, blockchain and reputation-based crowdsourcing iot federated learning: A secure, decentralized and privacy-preserving system," arXiv:1906.10893, 2019.
[19] The TFF Authors. (2019). Tensor Flow Federated. Accessed: Sep. 20, 2021.
[20] The Leaf Authors. (2019). Leaf. Accessed: Sep. 20, 2021.
[21] Y. Chen, Y. Liu, J. Yang, Z. Li, & Y. Li, . Federated Learning for Speaker Verification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,2019
[22] X. Li, X. Huang, Z. Wang, W. Gao, & J. Chen, Adaptive Communication for Federated Learning in Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*,2020.
[23] H. Zheng, S. Zhang, Y. Wang, B. Hu & Y. Liu, Transfer Learning for Data Heterogeneity in Federated Speaker Verification. *International Conference on Pattern Recognition (ICPR)*,2021
[24] Q.Wang, S. Wu, H. Wang, L. Zhang, & X. Li, Federated Meta-Learning for Speaker Verification. *IEEE International Conference on Multimedia and Expo (ICME)*,2022.
[25] A. Gupta, "How Federated Learning is going to revolutionize AI," towards data science, [Online]. Available: https://towardsdatascience.com/how-federated-learning-is-going-to-revolutionize-ai6e0ab580420f.
[26] A.Brecko, E. Kajati, J. Koziorek, I. Zolotova, "Federated Learning for Edge Computing: A Survey". Appl. Sci. 2022, 12, 9124.
[27] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. Arcas, "Communication-efficient learning of deep networks from decentralized data". In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54, Ft. Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
[28] U.Bhattacharjee, and K.Sarmah, "A Multilingual Speech Database for Speaker Recognition," In Proc. IEEE, ISPCC, March 2012.
[29] U.Bhattacharjee, and K.Sarmah, "Development of a Speech Corpus for Speaker Verification Research in Multilingual Environment," International Journal of Soft Computing and Engineering , Volume-2, Issue-6, pp. 443-446, January 2013.