Personalized Federated Learning on NLOS Acoustic Signal Classification

Hucheng Wang[1], Suo Qiu[2], Jingjing Wang[3], Lei Zhang[4], Zhi Wang[2], and Xiaonan Luo[1]

[1]Guilin University of Electronic Technology
[2]Zhejiang University
[3]Kyungpook National University
[4]Chang'an University

February 15, 2023

Abstract

In the process of identifying non-line-of-sight (NLOS), acoustics-based indoor positioning needs to collect audio recordings of sound fields in multiple rooms and upload them to the central server for training. Once the transmission process and server-side suffer malicious attacks, private data will also be leaked. To solve the training difficulty and privacy issues at the same time, we propose a novel Personalized Federated Learning (PFL) model combined with user frequency and room data capacity, taking into account the significant differences in positioning data with room layout. The proposed model can accurately identify the differences between different room data when aggregating on the server-side. By collecting data in the actual indoor environment and comparing the existing algorithms, the accuracy of the proposed method in the data verification of unfamiliar rooms is 90%.

Personalized Federated Learning on NLOS Acoustic Signal Classification

Hucheng Wang, Suo Qiu, Jingjing Wang, Lei Zhang, Zhi Wang, Xiaonan Luo

***Abstract*—In the process of identifying non-line-of-sight (NLOS), acoustics-based indoor positioning needs to collect audio recordings of sound fields in multiple rooms and upload them to the central server for training. Once the transmission process and server-side suffer malicious attacks, private data will also be leaked. To solve the training difficulty and privacy issues at the same time, we propose a novel Personalized Federated Learning (PFL) model combined with user frequency and room data capacity, taking into account the significant differences in positioning data with room layout. The proposed model can accurately identify the differences between different room data when aggregating on the server-side. By collecting data in the actual indoor environment and comparing the existing algorithms, the accuracy of the proposed method in the data verification of unfamiliar rooms is 90%.**

***Index Terms*—Acoustic, NLOS, Federated Learning**

I. Introduction

Indoor positioning technology based on acoustic signal has the advantages of compatibility, stability, and high positioning accuracy. The outliers caused by NLOS are the most severe reason for the decline of positioning accuracy [1]. When identifying NLOS signals, the processing method based on the neural network first collects data in the form of signals, then applies the algorithms of the neural network to the collected data and finally achieves the purpose of identifying NLOS signals(e.g., [2], [3]).

The global training model cannot accurately predict the results of rooms with occlusion problems due to different layout types. The information data set related to the indoor environment is uniformly uploaded to the server. Once the malicious information attacker appears on the transmission path or server, the indoor privacy information will be completely exposed [4].

Federated learning (FL) [5] proposes a distributed learning architecture that allows each client to avoid uploading its original dataset to the server but to upload model parameters, like gradients. Meanwhile, FL conveys the idea of edge computing, enabling beacons in the room to respond to user data more quickly. FL not only prevents malicious attackers from directly obtaining the original user's privacy data but also greatly reduces the bandwidth pressure in the upload process.

However, the general FL aggregation model has requirements for the independence and distribution of data. For non-independent and identically (Non-IID) data generated by different room layouts, the global model often cannot adapt to the NLOS distribution of all rooms.

The private model based on each user node proposed by PFL [6] solves the problem of overfitting a single global model on Non-IID data. The federated averaging (FedAvg) [7] proposed fluctuates wildly and reduces the validation accuracy on some room datasets. Bui [8] updated FL training by embedding personalized parameters. Liang [9] proposed the LG-FedAvg algorithm that combines local representation learning and global federated training. However, due to the large difference in room layout, the LG-FedAvg cannot effectively solve the problem of verification failure. Based on the above problems, we propose our algorithm and compare and verify it with other methods.

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61773344, 61273079, 61772149, 61936002, and 6202780103.

II. FL aided NLOS Classification Design

The general indoor environment is considered to be composed of multiple different independent rooms, and the sound propagation model caused by different indoor layouts and occlusion distributions meets the Non-IID model. Assuming that the number of speakers in each room is S, the whole amount room is R, the audio received by the microphone $\mathbf{A}_t$ at time t in speaker s and room r can be expressed as

$$\mathbf{A}_t = \{\mathbf{A}_t^{1,1}, \mathbf{A}_t^{1,2}, ..., \mathbf{A}_t^{r,s}, ..., \mathbf{A}_t^{R,S}\}. \tag{1}$$

Each $\mathbf{A}_t^{r,s}$ is considered to be independent of each other, and reverberation only occurs between the LOS signal and the self-reflected signal.

A. Capture audio model

We assume that the indoor acoustic field consists only of direct sound waves, first-reflected and diffuse reflection waves, ignoring the weak multiple reflections and other related waves. The signal-to-noise ratio (SNR) of the transmit power is limited to p_{snr}, which causes the propagation distance of a single speaker to be limited. Due to the propagation speed c of sound waves, the sound cycle T_c of each speaker only needs to satisfy $T_c > S \cdot d_{max}/c$ to conform to $\mathbf{A}_t^r$, where multiplying by the number of speakers S means that the next round of sound can be broadcast only after all speaker signals are accepted. The distance between adjacent speakers shall not exceed d_{max}.

Considering the complex indoor environment, the captured audio signal $\mathbf{A}_t^{r,s}$ has three kinds of mixtures: LOS, reflection and diffusion signal, which can be described as [10]:

$$\mathbf{A}_t^{r,s} = \begin{cases} \alpha_L(\mathbf{S}_t^{r,s}) * w(t), & LOS, \\ [\sum_{m=1}^{N_{Re}} \alpha_{Re}^m(\mathbf{S}_t^{r,s}) + \sum_{n=1}^{N_D} \alpha_D^n(\mathbf{S}_t^{r,s})] * w(t), & NLOS, \end{cases} \tag{2}$$

where $\mathbf{S}_t^{r,s}$ denotes the raw acoustic signal. To simplify the system, the raw signal $\mathbf{S}$ is the same at all R rooms and S speakers. α_L, α_{Re}, and α_D are the attenuation of LOS, reflection and diffusion, respectively. The superscript m and n are m^{th} and n^{th} path in reflection and diffusion wave. The Blackman window $w(t)$ is used to filter out the low SNR signals that cannot be perceived.

B. Training model

The real-time positioning system needs to instantaneous exchange data from the user to the server. The user to be positioned sends the collected acoustic signal $\mathbf{A}_t^{r,s}$ to the server, and the server processes to determine the sight status and calculate the coordinates of the user. Further, judging the state of sight is done by the CNN and Bi-LSTM models of the deep neural network architecture. The short-term Fourier transform (STFT) is one of the most effective feature extraction methods left. After STFT, the complex value of the spectrum matrix $\mathbf{A}_t^{r,s}$ at each time t is obtained as the input layer of CNN. CNN extracts multi-dimensional features by convolution and pooling layers iteratively and passes them to the Bi-LSTM network for classification training. Then, the probability of the current state is calculated by the fully connected layer, and the result of the NLOS state judgment is obtained. This process can be expressed as

$$\phi := \mathbf{C}_t \to \{1, 0\}, \mathbf{C}_t \in \mathbb{C}^{F \times T}, \tag{3}$$

$$\mathbf{C}_t = \sum_{r \in R} \sum_{s \in S} (STFT(\mathbf{A}_t^{r,s}) + \mathbf{N}), \mathbf{C}_t \in \mathbb{C}^{F \times T}, \tag{4}$$

where $\mathbf{N}$ is additional white noise. F and T are the dimension of the spectrum matrix $\mathbf{U}_t^{r,s}$, where F is the amount of frequency segment, and T is the amount of time segment in STFT, decided by window length and overlap length. In fact, ϕ is only a binary classification problem, while 0 means NLOS and 1 means LOS results.

C. PFL in NLOS classification

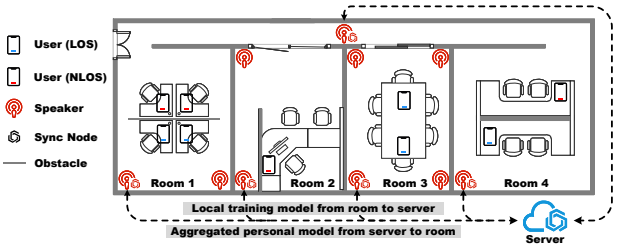

Fig. 1. Schematic diagram of room to server

The federated learning model protects data privacy by transmitting gradients or weights instead of training data. The server weights and aggregates the model of each client into the personal server model and then distributes it to each room. The room combines the self-training state with the personal server model, updates its training model, and completes an update process from the server to each room. Due to the different indoor layouts in each room and different occlusion scenarios, the speaker arrangement's Geometric Dilution Precision (GDOP) is also different, which is a typical Non-IID and identically distributed model. The specific steps are described as follows and the diagram as Figure 1.

1) Room model training: The same method is used to train each room separately, the model of r^{th} rooms can be described as

$$\phi^{(r)} := \mathbf{R}_t \to \{1, 0\}, \mathbf{R}_t \in \mathbb{C}^{F \times T}, \tag{5}$$

$$\mathbf{R}_t = \sum_{s \in S} (STFT(\mathbf{A}_t^{r,s}) + \mathbf{N}), \mathbf{R}_t \in \mathbb{C}^{F \times T}. \tag{6}$$

Then model set of each room $\phi = \{\phi^{(1)}, \phi^{(2)}, ..., \phi^{(R)}\}$ and the gradient $\nabla\phi$ from local SGD will be sent to server.

2) Server Aggregation: The accuracy of individual clients may decrease when the data distribution of each participant in federated learning is inconsistent. The cloud server side corresponds to the private model $\phi'^{(r)}$ of each room $\phi^{(r)}$, and the total server model $\phi' = \{\phi'^{(1)}, ..., \phi'^{(R)}\}$. To enable models with similar NLOS distributions to better adapt to training, instead of using a global model, the distance $||\phi_{E-1}^{(r)} - \phi_{E-1}^{(q)}||^2$ between $\phi^{(i)}$ and $\phi^{(j)}$ becomes important for the global update of the r^{th} room in the server. The amount of data is also an important indicator of aggregation. We allow each room r to perform E epochs of local room model update via mini-batch SGD with the size of n_r, then $\sum_{r \in R} n_r$ is the whole E training batches. The update process of the server personal model $\phi'^{(r)}$ corresponding to room r is

$$\begin{aligned} \phi_E'^{(r)} =& (1 - \alpha_E \sum_{q \neq r}^{R} ||\phi_{E-1}^{(r)} - \phi_{E-1}^{(q)}||^2) \frac{n_r}{\sum_{r \in R} n_r} \phi_{E-1}^{(r)} \\ &+ \alpha_E \sum_{q \neq r}^{R} ||\phi_{E-1}^{(r)} - \phi_{E-1}^{(q)}||^2 \frac{n_q}{\sum_{r \in R} n_r} \phi_{E-1}^{(q)} \\ &= \xi^{(r,1)} \phi_{E-1}^{(1)} + \cdots + \xi^{(r,R)} \phi_{E-1}^{(R)}, \end{aligned} \tag{7}$$

where $\xi^{(r,1)}, ..., \xi^{(r,R)}$ are the linear combination weights of the model parameter sets $\phi_{E-1}^{(1)}, ..., \phi_{E-1}^{(R)}$, respectively. $\phi_E'^{(r)}$ is actually a convex combination of model sets, where $\xi^{(r,1)} + \cdots + \xi^{(r,R)} = 1$. $\phi_{E-1}^{(1)}, ..., \phi_{E-1}^{(R)}$ are individual models of each room. α_E is normalization factor in E epoch.

The aggregation progress Equation 7 reflects the weight of n_r in the server update. If there is no user in r^{th} room in the E epoch, no data is generated, and the room of personal parameter $\xi^{(r,r)} = 0$. This operation does not affect the update of other individual rooms model.

3) Room model updating: To further reduce computing in each room, we select the closest personal model of the server to the room model:

$$\phi_E^{(r)} = \arg\min_{\phi \in \phi_E'} (||\phi - \phi_{E-1}^{(r)}||^2). \tag{8}$$

Here, the process of E epoch from room model training to room model updating is completed.

III. Experiments and Results

We set $R = 4$ and speakers in each room $S = 4$. The four different rooms are Lab 1, Lab 2, Office 1, and Office 2. The speakers are distributed on the ceiling of the corners in each room to fully cover the sound field. Each speaker sends 800 acoustic signals, 400 of which are captured by the microphone in the LOS range, and the other 400 signals are in the NLOS range. Each data is 16-bit sampling rate, 1-second duration, .wav lossless format, and a total of 15Gb of audio data. The dataset is uploaded to IEEE Dataport [11].

A. Results on general training

To evaluate the performance, we employ the centralized training model, global model (FedAvg) [7] and the personal model (proposed) on the existing room. To balance the total

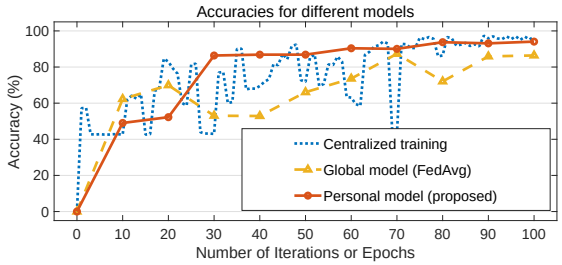

Fig. 2. Comparisons between different training of the existing room.

training epochs, we set the number of local training times multiplied by the updates times $E = 10$ is equal to 100 epochs. Although the number of centralized training epochs is up to 100, the accuracy constantly fluctuates, and it does not stabilize until after 80 epochs. For Non-IID room acoustic signals, FedAvg does not consider the difference in NLOS distribution in each room, causing the unstable validation accuracy. The proposed aggregation method takes into account the difference of the room. In the early validation stage, due to insufficient personal model training in each room, the accuracy is not ideal. Starting from the 30 epochs, the proposed method is significantly better than other methods, and there are no repeated fluctuations in the validation.

B. Results on unfamiliar room sound field

To verify the adaptability of the existing model to unfamiliar rooms, we remove the training data of a specific room. Then the removed room data is used to verify the trained model and calculated the NLOS classification accuracy of each model for the unfamiliar room.

As Figure 3 shows, the x-axis is the room that was removed in training. The PFL model showed the best performance when testing unfamiliar rooms. The method we proposed takes into account the room similarity distance $||\phi_{E-1}^{(r)} - \phi_{E-1}^{(q)}||^2$, and the sound field data of unfamiliar room will first select a similar personality model $\phi_E^{\prime(r)}$. Other methods do not take

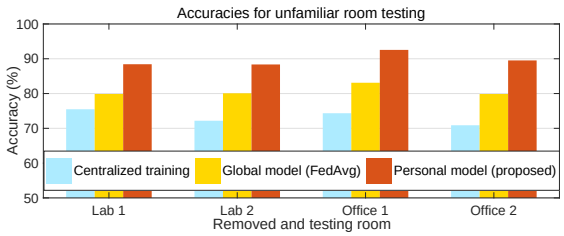

Fig. 3. Comparisons between different training of unfamiliar room.

individual considerations into account, and large outliers may be generated when the data in an unfamiliar room does not conform to the global model data distribution.

IV. Conclusion

This paper presented a personalized federated learning method on NLOS acoustic signal classification in indoor positioning. We are the first research to use the federated learning mechanism to NLOS classification practice indoors. We first conduct slimmable training on each local client terminal (mobile phone) to obtain a local model. Then, the server aggregates into a PFL model according to the model parameter distance and data volume and broadcasts it to the user for updates. Experiments have proved that the proposed method can make the general training accuracy not inferior to the centralized model and without training fluctuation. Significantly, the result in an unfamiliar room shows optimal performance.

References

[1] G. Wang, A. M.-C. So, and Y. Li, "Robust convex approximation methods for tdoa-based localization under nlos conditions," *IEEE Transactions on Signal processing*, vol. 64, no. 13, pp. 3281–3296, 2016.

[2] X. Ye, C. Ma, W. Liu, and F. Wang, "Robust real-time kinematic positioning method based on nlos detection and multipath elimination in gnss challenged environments," *Electronics Letters*, vol. 56, no. 24, pp. 1332–1335, 2020.

[3] S. Fan, Y. Wu, C. Han, and X. Wang, "A structured bidirectional lstm deep learning method for 3d terahertz indoor localization," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 2381–2390.

[4] A. K. Singh and J. Kumar, "Secure and energy aware load balancing framework for cloud data centre networks," *Electronics Letters*, vol. 55, no. 9, pp. 540–541, 2019.

[5] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[6] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *arXiv preprint arXiv:2103.00710*, 2021.

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

[8] D. Bui, K. Malik, J. Goetz, H. Liu, S. Moon, A. Kumar, and K. G. Shin, "Federated user representation learning," *arXiv preprint arXiv:1909.12535*, 2019.

[9] P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," *arXiv preprint arXiv:2001.01523*, 2020.

[10] L. Zhang, M. Chen, X. Wang, and Z. Wang, "Toa estimation of chirp signal in dense multipath environment for low-cost acoustic ranging," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 2, pp. 355–367, 2018.

[11] H. Wang, S. Qiu, and L. Zhang, "Nlos/los acoustic signal dataset," 2021. [Online]. Available: https://dx.doi.org/10.21227/en06-cn42