Expert Systems With Applications 287 (2025) 128154

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Federal parameter-efficient fine-tuning for speech emotion recognition

Haijiao Chen [a], Huan Zhao [a,*], Zixing Zhang [a,*], Keqin Li [a,b]

[a] College of Information Science and Engineering, Hunan University, Changsha, 410082, China
[b] Department of Computer Science, State University of New York, New Paltz, New York, 12561, USA

ARTICLE INFO

Keywords:
Anti-attribute inference attacks
Cloud-edge-terminal
Federal learning parameter-efficient fine-tuning
Pre-trained speech model
Privacy preservation
Speech emotion recognition

ABSTRACT

Pre-trained speech models leverage large-scale self-supervised learning to create general speech representations, with fine-tuning on specific tasks like Speech Emotion Recognition (SER) significantly enhancing performance. However, fine-tuning on different datasets necessitates storing full copies of model weights, leading to substantial storage demands and deployment challenges, particularly on resource-constrained devices. Centralized training also poses substantial privacy risks due to direct access to raw data. To address these challenges, we propose a cloud-edge-terminal collaborative paradigm for Federal Learning Parameter-Efficient Fine-Tuning (FedLPEFT), which harnesses the synergy of cloud and edge computing to drive the development of collaborative SER applications. Specifically, the distributed paradigm of Federated Learning (FL) offers a privacy-preserving schema for collaborative training, and fine-tuning based on pre-trained speech models can improve SER performance. Parameter-Efficient Fine-Tuning (PEFT) embeds trainable layers in the feed-forward layers of pre-trained speech models. By freezing backbone parameters and sharing only a small set of trainable parameters, PEFT reduces communication overhead and enables lightweight interactions. Additionally, our experiments on attribute inference attacks across various pre-trained models show that gender prediction is at chance levels, indicating that the FedLPEFT approach significantly mitigates sensitive information leakage, ensuring robust privacy protection.

1. Introduction

Speech emotion recognition (SER) identifies a speaker's emotional state by analyzing the acoustic features of audio signals, enhancing human-computer interaction and supporting emotional needs. Accurate emotion detection also provides valuable insights for emotionally responsive smart products. Despite the significant challenges posed by the complex processing logic of speech signals and performance bottlenecks, SER continues to attract research interest. It is widely used in fields such as autonomous driving, voice assistants, smart homes, conversational systems, health monitoring, and intelligent education (Dixit & Satapathy, 2024; Kim & Hong, 2024; López-Gil & Garay-Vitoria, 2024). Recent advancements in deep learning, particularly Transformer-based methods (Chen et al., 2023a; Khan, Gueaieb, El Saddik, & Kwon, 2024), have shown notable performance in SER. However, these methods rely on self-attention mechanisms to process long sequences, enabling the model to attend to all input elements and capture long-range dependencies, thereby substantially increasing computational complexity. The advent of large conversational models like ChatGPT has intensified research into pre-trained models. Pre-trained speech models offer a novel approach for SER, generating general speech representations through self-supervised or weakly supervised learning on large-scale data, and transferring this knowledge to downstream tasks for promising results (Gao, Zhou, Liu, Zhao, & Wen, 2023). However, fine-tuning on different datasets for various tasks requires saving full copies of the pre-trained speech model's weight parameters, making deployment on resource-constrained devices infeasible. Additionally, fine-tuning pre-trained models under centralized conditions relies on user data, posing challenges to data security and model privacy on terminal devices.

The distributed privacy-preserving paradigm of federated learning (FL) (Zhang et al., 2021) enables multiple parties to collaborate on model training while keeping data local and private through parameter-sharing mechanisms. This facilitates the development of a FL-based pre-trained model fine-tuning approach. However, practical implementation is challenged by the vast number of parameters in pre-trained models (often in the trillions) as well as frequent parameter exchanges, limited computational and bandwidth resources, and long communication links. Considering the growing demand for such collaborative services, we propose to build a cloud-edge-terminal collaboration architecture that integrates cloud and edge computing strategies to leverage their respective strengths. Cloud computing resolves limitations in device computation, communication, and storage, while edge computing alleviates issues with bandwidth and communication latency by deploying edge servers closer to users. As shown in Fig. 1, given that the computing

* Corresponding authors.
E-mail addresses: chenhaijiao@hnu.edu.cn (H. Chen), hzhao@hnu.edu.cn (H. Zhao), zixingzhang@hnu.edu.cn (Z. Zhang), lik@newpaltz.edu (K. Li).

https://doi.org/10.1016/j.eswa.2025.128154
Received 1 October 2024; Received in revised form 26 February 2025; Accepted 11 May 2025
Available online 15 May 2025

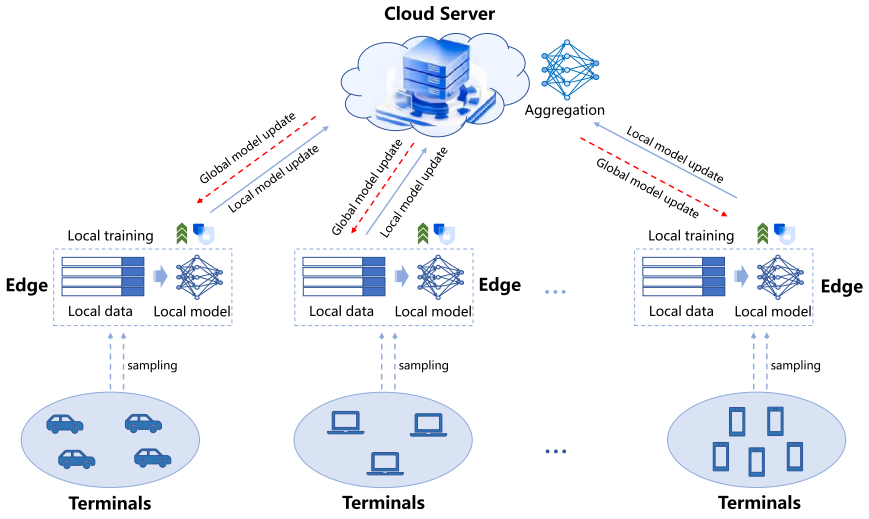

Fig. 1. The cloud-edge-terminal collaboration framework based on federated learning.

and storage capabilities of edge servers are much higher than those of terminal devices, we suggest placing FL logic between edge and cloud servers in this setup: edge servers manage data collection, storage, and local model training, cloud servers handle parameters aggregation and large-scale computations, and terminal devices focus on encrypted data transmission to edge servers.

Despite the advantages of the cloud-edge-terminal architecture for FL, the massive parameter sizes of pre-trained models limit its broader adoption. Recent studies (Zhu, Liu, & Han, 2019) reveal that FL frameworks are vulnerable to attribute inference attacks, where attackers can deduce sensitive information (e.g., gender) from local parameter updates. Existing privacy-enhancing techniques for FL, such as secure multi-party computation (Agrawal, Shahin Shamsabadi, Kusner, & Gascón, 2019), homomorphic encryption (Gong et al., 2024), and differential privacy (Feng, Peri, & Narayanan, 2022), are computationally intensive. Adding noise to large pre-trained models is often inefficient, and differential privacy may fail to provide adequate protection when attackers have access to frequent model updates. To address these issues, we propose a Federal Learning Parameter-Efficient Fine-Tuning (FedLPEFT) approach. This method involves freezing most backbone model parameters and updating only a small number of trainable layers during federated training. This approach significantly reduces parameter volume and communication overhead, making federated fine-tuning practical while enhancing system performance and privacy protection. We evaluate the effectiveness of Parameter-Efficient Fine-Tuning (PEFT) techniques, such as adapter tuning, embedding prompt tuning, and low-rank adaptation (LoRA), along with downstream fine-tuning across various pre-trained speech models, and conduct simulations of attribute inference attacks to validate the approach's robustness and security. In summary, our work leverages pre-trained models' extensive knowledge for FL, with PEFT supporting low communication and robust privacy protection. The main contributions are as follows.

- A privacy-enhanced federal parameter-efficient fine-tuning for speech signal processing task is proposed, integrating the benefits of cloud and edge computing to extend its applicability to cloud-edge-terminal scenarios, thus optimizing computational efficiency for speech emotion recognition.
- We evaluate adapter tuning, embedding prompt tuning, and LoRA on five pre-trained speech models, comparing them with downstream fine-tuning. Results confirm that LoRA exhibits stable and superior performance across all pre-trained systems. This fine-tuning method, which involves freezing the backbone parameters of pre-trained models, effectively reduces communication overhead while maintaining performance in federated interactions.
- We perpetuate an attribute inference attacks framework for federated learning pre-trained speech model fine-tuning scenarios and validate its effectiveness in enhancing privacy. The involvement of a limited number of trainable parameters in collaborative sharing reduces attribute inference capabilities to a chance level.

The remainder of this article is organized as follows. In Section 2, we review related work. Section 3 provides a detailed discussion of FedLPEFT methods based on various pre-trained speech models, and outlines the attack process. Section 4 analyzes the framework's performance, attack resilience, and parameter sensitivity, and compares it with centralized fine-tuning and related approaches. Finally, we summarize in Section 5.

2. Related work

Advancements in artificial intelligence (AI) and human-computer interaction have driven significant interest in SER research. Recently, deep learning-based approaches (Latif et al., 2021), particularly those utilizing transformers (Vaswani et al., 2017) and their variants, have shown exceptional performance. These models' advanced attention mechanisms allow for the effective capture of critical semantic information in long sequence datasets by simultaneously addressing both global and local features, thereby improving system performance and generalization. Wang et al. (2021) developed an end-to-end SER architecture by stacking multiple transformer layers to enhance global feature aggregation, achieving a 20 % performance improvement. Chen, Lin, Wang, Zheng, and Liu (2023b) introduced a spatio-temporal representation learning approach with a multi-head attention mechanism, combining fine-grained frame-level and coarse-grained utterance-level emotional features to boost SER performance. Naderi and Nasersharif (2023) utilized Wav2Vec2.0 transformer blocks and prosodic features, effectively extending SER to cross-corpus tasks using fused attention and transfer learning techniques. Additionally, due to the natural graph structure of conversations, many speech-based conversational emotion recognition studies leveraging Graph Convolutional Networks (GCN) have yielded significant outcomes. Chandola, Altarawneh, Jenkin, and Papagelis (2024) proposed a two-stage method to predict speaker emotions by first extracting utterance-level features and then using these to form dialogue graphs for GCN training. Yuan et al. (2023) introduced a relational dual-layer aggregation GCN, aimed at reducing redundancy and preserving node information during aggregation.

Research on large-scale pre-trained speech models is increasingly dominant, driven by the models' ability to acquire general knowledge through self-supervised and weakly-supervised learning on extensive

datasets. Zhang et al. (2024) comprehensively analyzed the significant contributions of large language models (LLMs) to SER tasks, advocating for broader discussions on enhancing speech emotion recognition with more advanced and generalized models. However, fine-tuning large pre-trained models requires storing entire weight parameter copies, which imposes substantial storage constraints, limiting practical deployment. To address this, PEFT methods have been introduced, inserting trainable layers into pre-trained models and optimizing only a small subset of parameters. Studies like (Ding et al., 2023; Li et al., 2023) have thoroughly evaluated the benchmarks of PEFT on pre-trained models. Lashkarashvili, Wu, Sun, and Woodland (2024) systematically investigated the effectiveness of various PEFT methods for both discrete emotion classification and dimensional emotion attribute prediction. Li and Hou (2023) combined self-supervised learning features with adapter fine-tuning for SER, demonstrating that adapter fine-tuning significantly enhances the transferability of self-supervised learning features across different tasks, offering new insights into the advancement of speech processing. Despite these advancements, the critical issues of data security and model privacy in pre-trained speech models remain underexplored-key challenges for the sustainable development of AI.

Centralized training exposes significant data security risks due to direct access to user data, while fine-tuning pre-trained models is vulnerable to parameter leakage. The advent of FL offers a potential solution by integrating privacy-preserving frameworks with pre-trained models. However, recent studies (Feng, Hashemi, Hebbar, Annavaram, & Narayanan, 2021; Zhao, Chen, Xiao, & Zhang, 2023a) have highlighted the privacy vulnerabilities of deep models within FL frameworks, raising concerns about their effectiveness in safeguarding privacy. Traditional encryption methods, such as homomorphic encryption, secure multi-party computation, and differential privacy, are impractical for large-scale pre-trained models due to the additional computational burden they impose. Some adversarial training techniques, like those proposed by Ren, Baird, Han, Zhang, and Schuller (2020), use methods such as Fast Gradient Sign Method (FGSM) to generate adversarial data for training defenses. Jaiswal and Provost (2020) employ adversarial learning to capture private, overlooked information in multi-model representations. However, these methods may fail as attackers can still access the perturbed data. Chen, Zhao, Zhang, and Li (2024) proposed a federated lightweight distillation approach with parameterized hierarchical sharing to enhance privacy in federated pre-trained models, but its extensive preprocessing requirements may limit its applicability.

Several works combining parameter-efficient fine-tuning with federated learning (Malaviya, Shukla, & Lodha, 2023; Sun, Li, Li, & Ding, 2024; Zhang et al., 2023; Zhao, Du, Li, Li, & Liu, 2023b) have been explored in NLP, primarily using text datasets. However, our focus on audio inputs (e.g., raw waveforms, spectrograms) and speech models introduces distinct challenges due to domain differences. Moreover, speech data contains rich paralinguistic cues, making sensitive attribute leakage (e.g., gender, age) a critical threat for attribute inference attacks. Therefore, the FedLPEFT approach for pre-trained speech models remains underexplored and needs thorough evaluation regarding SER performance, parameter privacy, and communication overhead.

3. Proposed approach

To tackle federated privacy leakage in cloud-edge-terminal scenarios, we propose a PEFT paradigm designed for cloud-edge-end architectures to enhance federated privacy. Our goals are: (1) to develop a federated learning-based PEFT framework for extending speech emotion recognition scenarios requirements; (2) to demonstrate the feasibility of combining federated learning with pre-trained speech models through PEFT, showing promising performance and providing guidance for federated large speech model development; (3) to validate robust privacy protection by simulating attribute inference attacks across multiple datasets, considering the particularity of speech data containing rich paralinguistic information. Fig. 2 illustrates the federated privacy protection system framework, comprising two main components. The upper part details the core logic of federated privacy enhancement, divided into private training and shadow training. Private training handles SER tasks, while shadow training simulates privacy training with different public datasets for attribute inference (e.g., gender prediction). Our local models utilize pre-trained speech models with various PEFT methods, such as adapter tuning, embedding prompt, LoRA, and downstream fine-tuning. The lower part provides details on attribute inference attacks. Further specifics are discussed in the following sections.

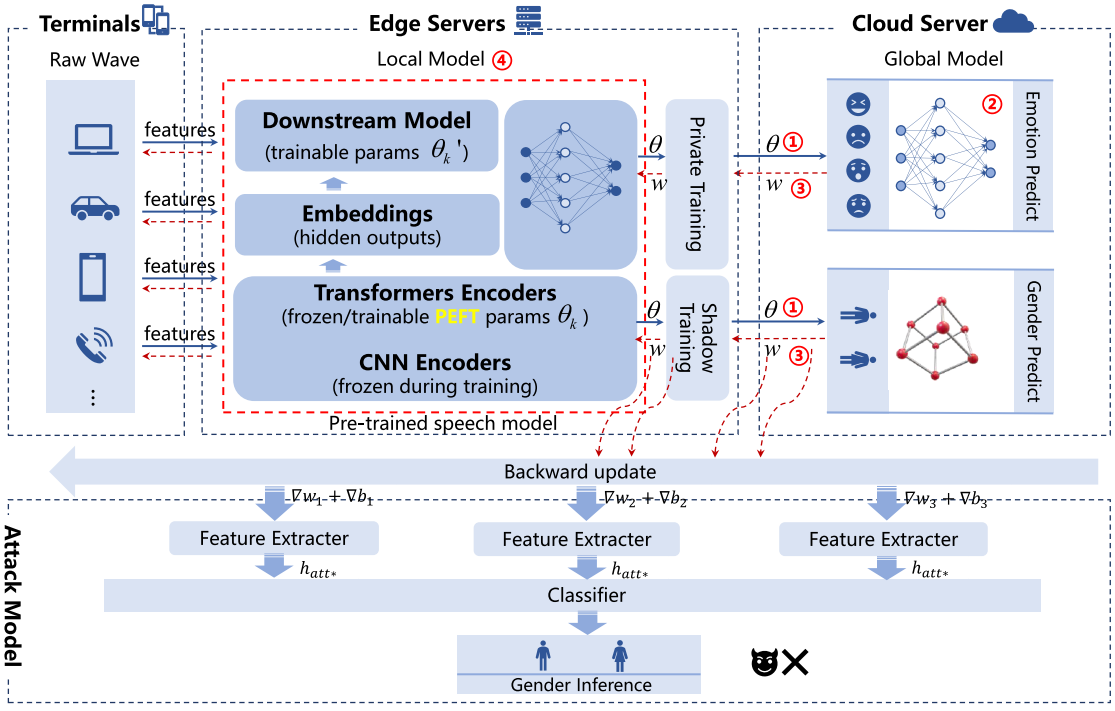

Fig. 2. Federal Learning Parameter-Efficient Fine-Tuning (FedLPEFT) framework based on pre-trained speech models enhances privacy protection.

3.1. PEFT privacy-enhanced paradigms in FL

Pre-trained speech models have enhanced SER performance due to large-scale data training. However, in FL settings, they present challenges such as increased communication burdens from a large number of model parameters and potential privacy leaks. To address these issues and explore the feasibility of integrating large pre-trained speech models with FL for SER tasks, the FedLPEFT privacy-enhanced algorithm has been proposed. Algorithm 1 details the process, highlighting key steps in the FL collaborative training,

Algorithm 1 FedLPEFT for Privacy Preservation.

Input: Pre-trained model M, fine-tuning method F, pre-trained model trainable parameters θ_k, downstream model trainable parameters $\theta_k{}'$, local model w_k, local model updates θ, global model w_g, global model updates w, learning rate η, local epoch α, num epochs β, training round t, learning step τ;

Output: SER global test accuracy (ACC), Unweighted Average Recall (UAR), best model;

Initialize global model w_g and broadcast it and related parameters to the selected k active client, $w_k^{t,0} \leftarrow w_g$;
repeat
  for selected k clients in parallel **do**
    for fold = 1, 2, …, 5 **do**
      for t = 1, 2, …, β **do**
        Train local models and minimize loss, trainable layers involved,
        $w_k^{t^*} = \arg\min_{w^t} L(w_k^t)$;
        Update the local model,
        $w_k^{t,\tau+1} \leftarrow w^{t,\tau} - \eta w_k^{t^*,\tau}$;
        Upload local model parameters θ to the server,
        $\theta = \theta_k + \theta_k{}'$;
        Server aggregate and average parameters, and minimize global loss, in tth round training, K out of N clients participated, sample size n_k at the k client,
        $w^{t+1} = \sum_{k=1}^{K} \frac{n_k}{N} \theta_k^t$, $L(w_g^t) = \frac{1}{N} \sum_{k=1}^{N} L(w^t)$;
      end for
      Save the best model and results;
    end for
    Calculate the average of the 5-fold experiments;
  end for
until training stop or converge

a) Global model synchronization. Initialize the global model. The FL server randomly selects k active clients and broadcasts the latest global model w_g and parameters to them.

b) Local training and updates. Each client trains on its local dataset and updates its local model. With PEFT (adapter tuning/embedding prompt tuning/LoRA) and downstream model fine-tuning, most backbone parameters are frozen, and only trainable parameters (*requires grad = True*) are updated.

c) Model aggregation. Aggregate and average the local parameters from clients. The averaged global model parameters are then sent back to the clients. The updated local model combines the server's global model (trainable layers) with the client's local model (frozen layers), repeating until convergence.

We detail several PEFT methods, including adapter tuning, embedding prompt, and LoRA, as shown in Fig. 3. All fine-tuning methods are based on the FL pre-trained speech model. Before discussing these methods, we briefly summarize the pre-trained speech models used in this work.

3.1.1. Pre-trained speech model

Wav2vec 2.0 (Baevski, Zhou, Mohamed, & Auli, 2020) is a transformer-based self-supervised model that extracts contextual representations from raw audio. Its core components include a feature encoder, transformer module, and quantization module, trained with a masked learning objective to predict quantized speech representations.

WavLM (Chen et al., 2022), built on the Hubert framework, uses a denoising masked speech modeling approach. It enhances model understanding by masking parts of speech data and predicting the masked portions. Pre-trained on 94,000 hours of english speech, it excels in various speech-based tasks.

Whisper (Radford et al., 2023) is a transformer-based model trained with weak supervision. Designed for versatility, it handles speech input across different languages, dialects, and noise environments. The series includes versions like Whisper Tiny, Base, Small, and Large, each varying in parameters, computational needs, and accuracy.

3.1.2. Adapter tuning

Referring to Fig. 3(a), the fine-tuning technique adapts large-scale pre-trained models by inserting lightweight adapter modules between certain layers, allowing the model to adapt to new tasks while keeping most of the original parameters frozen. As described in Houlsby et al. (2019), the adapter adds trainable layers to each transformer layer, typically including a down-projection layer, an up-projection layer, and a nonlinear activation function. The down-projection reduces the dimensionality of the transformer's output, projecting high-dimensional

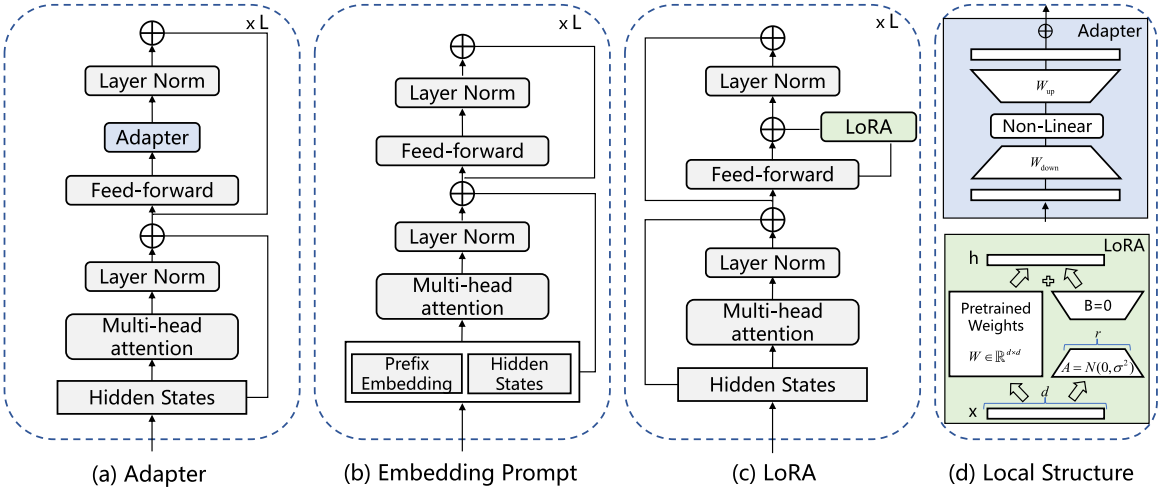

Fig. 3. Model structures and internal details from various PEFT methods.

features h (dimension d) into a lower-dimensional space m ($m \ll d$). The reduced features are then activated using a nonlinear function σ (e.g., ReLU) and projected back to the original dimensions. The adapter's output is then added to the original input features, producing the adjusted output. A skip-connection ensures the model can revert to an identity function if necessary. The adapter's computation is given by:

$$h = h + \sigma(hW_{\text{down}})W_{\text{up}}. \tag{1}$$

With the majority of parameters frozen and only a minimal set within the adapter being trained, this approach optimizes the FL framework by reducing computational resource demands and parameter storage requirements.

3.1.3. Embedding prompt tuning

Fig. 3(b) illustrates that a trainable embedding prompt is inserted into the input embedding space preceding each encoder in the pre-trained model. Embedding prompt fine-tuning, initially developed as a text prompting technique (Jia et al., 2022), updates these prompts during training to direct the model in generating task-specific outputs more effectively. The prompt output from the preceding layer is discarded and replaced with a new set of prompts before being fed into the subsequent layer. For downstream tasks, only the embedding prompt and classification layer parameters are optimized, while the core parameters of the pre-trained model backbone remain fixed. Given the original input H_i of the encoder's ith layer and the embedding prompt E_i, the concatenated input for the encoder during embedding prompt fine-tuning is:

$$H_i' = Encoder(concat(H_i, E_i)). \tag{2}$$

Indeed, this represents a lightweight and efficient fine-tuning approach, allowing large-scale pre-trained models to adapt effectively to diverse downstream tasks.

3.1.4. LoRA

A technique for efficient fine-tuning of large-scale language models involves introducing low-rank decomposition methods to reduce the number of parameter updates during fine-tuning, thereby lowering computational and storage costs while maintaining strong model performance (Hu et al., 2022). In an FL setting, LoRA is naturally suitable and advantageous; by decomposing the original parameter matrix into the product of two low-rank matrices, it can mitigate privacy concerns to some extent. Specifically, we keep the pre-trained layers fixed and apply trainable low-rank matrices to the feed-forward layers. Given a dense neural network layer with a weight matrix $W_0 \in \mathbb{R}^{d\times k}$, updating it with $\Delta W \in \mathbb{R}^{d\times k}$ results in an updated layer parameterized by $W = W_0 + \Delta W$. To improve computational efficiency, LoRA decomposes ΔW into two smaller matrices, $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, so $\Delta W = BA$, where $r \ll \min\{d, k\}$. During fine-tuning, the original weight matrix W_0 remains unchanged, and only A and B are updated. For an input feature h, the linear transformation adjusted by LoRA can be represented as:

$$h' = W_0 h + \Delta W h = W_0 h + (BA)h. \tag{3}$$

3.1.5. Downstream fine-tuning

Beyond the PEFT methods described, we establish a baseline fine-tuning method, downstream model fine-tuning, where the backbone network remains frozen and only the downstream model is fine-tuned. The downstream model architecture will be detailed in the experimental setup section. All fine-tuning methods use the same downstream classification structure. PEFT methods apply adapter tuning, embedding prompt tuning, and LoRA to each pre-trained architecture to assess SER performance across different pre-trained models and fine-tuning techniques.

3.2. Attack design

In a FL setting, where the primary task is SER, we utilize a private labeled dataset D_p from multiple clients, each containing an audio dataset X_a, emotion labels Y_a, and gender labels Z_a. The attacker, without access to D_p, can use a public dataset D_ρ, similar in format and distribution but non-overlapping. This work examines white-box attacks, where the attacker has full knowledge of the model architecture and hyperparameters, including learning rate, batch size, and epoch. The attack scenario is defined within the FedAvg frameworks (Collins, Hassani, Mokhtari, & Shakkottai, 2022), where the attacker (curious server) attempts to infer the gender attribute Z_k of the kth client by exploiting the shared model parameters θ_k and global model parameters w. Further details are as follows:

Private training. In the SER training process, private training is conducted using the FedAvg algorithm with PEFT, as illustrated in Fig. 2. Private data remains on the clients and is inaccessible to the server. However, shared training updates (gradients or parameters) are potentially vulnerable to indirect access by attackers. In our FedLPEFT method, clients share a subset of parameters: trainable PEFT parameters θ_k and downstream model parameters $\theta_k{}'$. In each tth training round, K clients with n_k samples collaboratively train, submitting their parameters to the server, which aggregates and averages them, the average parameters w are as follows:

$$\begin{cases} w^{t+1} = \sum_{k}^{K} \frac{n_k}{N} \theta_k^t, \\ \theta_k^t = \theta_k + \theta_k{}'. \end{cases} \tag{4}$$

Shadow training. Originally proposed in membership inference attacks (Nasr, Shokri, & Houmansadr, 2019), shadow training is adapted for attribute inference attacks in a similar framework. Like private training, its goal is to predict emotion categories. The model updates generated are stored to provide data for training the attack model. Specifically, the attacker trains a shadow model $M_1, M_2, \ldots, M_k$ using a public dataset D_ρ, with each shadow training dataset matching the format and distribution of the private dataset but not overlapping. For example, using IEMOCAP as the private dataset, MSP-IMPROV and CREMA-D are used for shadow model training, and vice versa. In this white-box attack scenario, the shadow model shares the same architecture and hyperparameters as the private model, including batch size, learning rate, and global epoch.

Attribute inference attacks. We construct the attack training dataset D_α from shared model updates generated during shadow training. The shared model g_k^t for the kth client is defined, with the client's gender label set Z_k used as labels for g_k^t. We then train the attack model M_α using D_α for gender inference. The attacker can access only the global model parameters w and the updated parameters θ_k^t from the kth client, not the original gradients. We use a pseudo-gradient g'^t_k derived from Geng et al. (2021) as input to the attack model, assuming T local updates and a learning rate of η,

$$g'^t_k = \frac{1}{T\eta}(w - \theta_k^t). \tag{5}$$

Thus, we train the attack model with parameters ϑ to minimize the following cross-entropy loss function:

$$\min_{\vartheta} L(M_\alpha(g'^t_k; \vartheta), Z_k). \tag{6}$$

Moreover, the attack model, as outlined in Feng et al. (2021), integrates a CNN feature extractor and a classifier. Weight updates ∇w_i and bias updates ∇b_i for the ith layer are generated during shadow training and input into a three-layer CNN to compute hidden representations (see Fig. 2). These CNN features are flattened and concatenated with the biases, then passed to an MLP for gender prediction. Given that the first-layer updates typically contain more sensitive information and are more susceptible to leakage, we evaluate the attack model's performance using ∇w_1 and ∇b_1.

Table 1
Emotion label statistics across three different emotion datasets.

Datasets	Neutral	Happy	Sad	Angry	Total
IEMOCAP	1708	1636	1084	1103	5531
CREMA-D	1972	1219	588	1019	4798
MSP-IMPROV	3477	2644	855	792	7798

4. Experiment

4.1. Datasets

In this work, we develop FedLPEFT and attack methods for SER using three widely utilized public datasets. Given the imbalanced data distribution, we adopt the four emotion labels frequently employed in the literature, as recommended in Khan et al. (2024), Chen et al. (2023a), Li, Liu, Yang, Sun, and Wang (2021): neutral, happy, sad, and angry. Table 1 presents the label distribution across these datasets.

The IEMOCAP corpus (Busso et al., 2008), collected by the University of Southern California, comprises multimodal recordings, including motion, audio, and video, specifically designed to capture explicit emotional expressions. The dataset consists of 5531 utterances from 10 actors (5 male and 5 female).

The CREMA-D corpus (Cao et al., 2014) contains audiovisual multimodal recordings from 91 actors (48 male and 43 female) expressing emotions. It includes 4798 utterances categorized into four emotions: neutral, happy, sad, and angry.

The MSP-IMPROV corpus (Busso et al., 2016) is designed to investigate natural emotions captured in improvisational scenes. It includes audio and visual data recorded under improvisational, natural, target, and read conditions, with 12 participants (6 male and 6 female). To comprehensively assess PEFT performance across various scenarios, we utilized data from all recording conditions, comprising a total of 7798 utterances.

4.2. Data preprocessing

For each dataset, 20 % of speakers are reserved as the test set, with an 8:2 split. Due to the limited number of speakers in IEMOCAP (10 speakers) and MSP-IMPROV (12 speakers), these datasets are divided into 10 subsets by speakers to align with CREMA-D, which has 91 speakers. For the new speaker subsets, 80 % is used for training and 20 % for validation. We employ 5-fold cross-validation and report average results. In FL training, we use data from 20 % of the speakers, with 8 clients receiving balanced data from 1 to 3 non-overlapping speakers each. Data partitioning for centralized PEFT methods follows (Feng & Narayanan, 2023).

4.3. Experimental settings

Model setup and evaluation. This work designs a model for speech emotion recognition, consisting of a pre-trained model and a downstream model. Fig. 2 shows the pre-trained models: Wav2vec 2.0 Base, WavLM Base+, Whisper Small, Whisper Base, and Whisper Tiny. The core modules include CNN encoders (frozen during training) and transformer encoders (with only adapter/embedding prompt/LoRA parameters trainable), processing raw audio input. The downstream model follows the architecture in Mireshghallah et al. (2020), which combines CNN encoder outputs using weighted averaging, applies three 1D convolutional layers (256 filters, kernel size 1, ReLU activation), averages the convolutional outputs, and feeds them into two fully connected layers for SER prediction. The experiments compare adapter, embedding prompt, and LoRA PEFT methods with downstream fine-tuning, which keeps the backbone encoder frozen and only fine-tunes the trainable parts. We set the low-rank dimension to 8, the embedding prompt size to 5, and the adapter size to 128, with 5 shadow models. For FedAvg, the learning rate is 0.0005, local epochs 1, and total epochs 30. During attack model training, the optimizer's learning rate is 0.0001, and client-specific trainable parameters, including embedding prompts, low-rank matrices, and downstream model parameters, are updated each round. For details on the centralized PEFT method settings, please refer to reference Feng and Narayanan (2023). All experiments use 5-fold cross-validation, with evaluation metrics being global test accuracy (ACC) and Unweighted Average Recall (UAR). The experiments are conducted on a Linux server with 4 GeForce RTX 3090 GPUs and 64GB RAM.

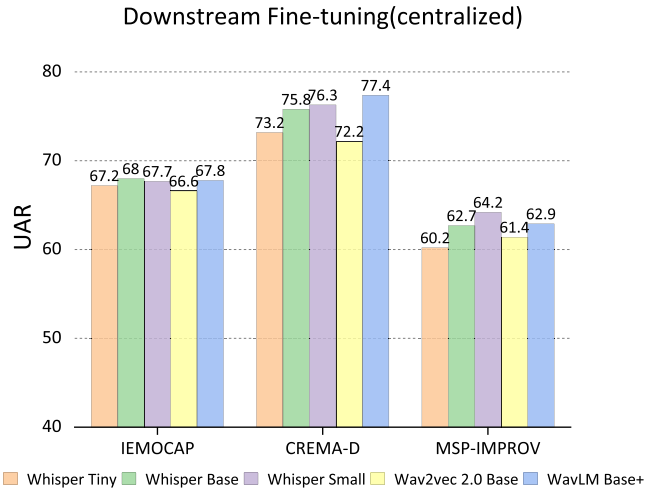

Fig. 4. Performance of downstream fine-tuning with various pre-trained models in centralized setups (UAR%).

4.4. SER performance

4.4.1. Centralized fine-tuning

(1) This section evaluates SER performance with different fine-tuning methods in a centralized environment. First, we give the specific performance of downstream model fine-tuning (UAR) as a baseline method. Fig. 4 shows that Whisper Small and WavLM Base+ deliver the best results across all datasets. Specifically, WavLM Base+ excels on the CREMA-D dataset, while Whisper Small outperforms other models on the MSP-IMPROV dataset. In contrast, Wav2vec 2.0 Base performs less competitively, even lagging behind the lighter models Whisper Base and Whisper Tiny. These results are consistent with those reported in Feng and Narayanan (2023).

(2) We compare three PEFT methods with downstream fine-tuning. Table 2 presents ACC/UAR metrics across three datasets, with average PEFT performance provided for clarity. The Average row shows that with Whisper models, adapter, embedding prompt, and LoRA perform worse than downstream fine-tuning. Conversely, PEFT methods enhance performance for Wav2vec 2.0 Base and WavLM Base+, with LoRA performing best. Among all models, WavLM Base+ consistently achieves the highest results, particularly with LoRA, which averages ACC: 71.5 % and UAR: 71.7 %. For individual datasets, Wav2vec 2.0 Base and WavLM Base+ show stable performance across fine-tuning methods. In contrast, Whisper models exhibit significant variability, with Adapter fine-tuning underperforming on Whisper Small and Whisper Base, and embedding prompt yielding the lowest results on Whisper Tiny, we speculate that the positional embeddings in Whisper contribute to this instability. Nevertheless, LoRA on Whisper models performs closest to downstream fine-tuning.

4.4.2. Federated fine-tuning

(1) In FL settings, parameter fine-tuning methods involve complex computations and interactions. We use frozen pre-trained model encoders and fine-tune downstream models to present baseline results.

Table 2
Performance comparison of various fine-tuning methods across pre-trained models in a centralized setting (ACC%/UAR%).

Dataset	Fine-Tune	Pre-trained Model									
		Whisper Tiny		Whisper Base		Whisper Small		Wav2vec 2.0		WavLM Base+	
		ACC	UAR	ACC	UAR	ACC	UAR	ACC	UAR	ACC	UAR
IEMOCAP	Downstream	**65.6**	**67.2**	**66.7**	**68.0**	**65.6**	**67.7**	64.4	66.6	65.7	67.8
	Adapter	62.8	63.9	57.9	58.1	56.5	59.4	59.6	63.8	66.9	69.1
	Prompt	55.2	57.5	57.5	59.9	59.0	59.5	**64.7**	**66.8**	68.4	70.2
	LoRA	63.7	65.8	65.4	67.3	63.1	65.8	63.6	66.3	**69.5**	**70.5**
CREMA-D	Downstream	**72.0**	**73.2**	74.1	75.8	73.1	76.3	69.9	72.2	77.1	77.4
	Adapter	70.0	71.4	69.6	70.8	62.5	64.8	69.3	71.5	76.8	77.6
	Prompt	62.0	65.0	69.0	71.5	70.4	72.8	73.0	73.5	77.2	78.9
	LoRA	71.9	72.1	**74.7**	**77.0**	**76.2**	**78.4**	**73.4**	**74.8**	**78.9**	**79.3**
MSP-IMPROV	Downstream	57.8	60.2	**62.8**	**62.7**	**63.4**	**64.2**	57.9	61.4	63.3	62.9
	Adapter	52.2	55.0	51.7	55.8	51.9	51.0	57.5	60.2	62.7	62.5
	Prompt	53.2	53.1	56.9	58.6	57.2	60.0	57.8	61.3	64.8	65.0
	LoRA	**58.5**	**61.1**	61.7	62.3	61.8	61.6	**58.8**	**61.4**	**66.0**	**65.4**
Average	Downstream	**65.1**	**66.9**	**67.9**	**68.8**	**67.4**	**69.4**	64.1	66.7	68.7	69.4
	Adapter	61.7	63.4	59.7	61.6	57.0	58.4	62.1	65.2	68.8	69.7
	Prompt	56.8	58.5	61.1	63.3	62.2	64.1	65.2	67.2	70.1	71.4
	LoRA	64.7	66.3	67.3	68.9	67.0	68.6	**65.3**	**67.5**	**71.5**	**71.7**

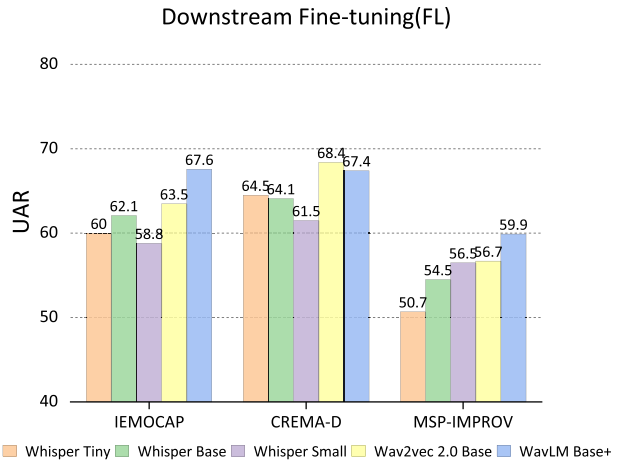

Fig. 5. Performance of downstream fine-tuning with various pre-trained models in FL settings (UAR%).

Fig. 5 shows that WavLM Base+ and Wav2vec 2.0 Base achieve the best performance across datasets. WavLM Base+ excels on IEMOCAP and MSP-IMPROV, while Wav2vec 2.0 Base leads on CREMA-D, with WavLM Base+ generally outperforming others. Wav2vec 2.0 Base's performance in federated learning contrasts with centralized findings. Whisper models underperform overall, with Whisper Tiny surprisingly outperforming Whisper Small on two datasets. These results indicate that model size and data distribution impact federated learning performance.

(2) We further evaluate PEFT methods across datasets, as shown in Fig. 6. WavLM Base+ and Wav2vec 2.0 Base demonstrate superior performance across all four FL fine-tuning methods. Compared to downstream model fine-tuning, adapter, embedding prompt, and LoRA deliver better results, enhancing the performance of these pre-trained models. This suggests that WavLM Base+ and Wav2vec 2.0 Base are well-adapted to federated learning environments, leveraging their extensive pre-training on large datasets to perform effectively on specific tasks. LoRA consistently outperforms other methods across pre-trained models, establishing it as the preferred approach in federated learning due to its reduced communication overhead. In contrast, adapter and embedding prompt methods exhibit suboptimal performance with Whisper models, falling significantly short of downstream fine-tuning and LoRA. This discrepancy is likely attributed to factors such as Whisper's positional embeddings, the FedAvg aggregation method, and insufficient learning of speech emotion features in the federated learning setting.

4.4.3. Centralized vs. FL distributed fine-tuning

To clearly compare FL and centralized settings across various pre-trained models and fine-tuning methods, Table 3 provides detailed insights. FL shows a slight performance drop compared to centralized methods, consistent with expectations due to FedAvg parameter averaging. In FL, Wav2vec2.0 Base and WavLM Base+ perform closer to centralized results, while most Whisper-based fine-tuning methods (except LoRA) underperform. LoRA stands out in the FL setup, especially when applied to the WavLM Base+ model, achieving the best results on the IEMOCAP (ACC: 69.2 % / UAR: 70.4 %), CREMA-D (ACC: 78.2 % / UAR: 75.0 %), and MSP-IMPROV (ACC: 65.1 % / UAR: 61.6 %) datasets, with only about a 4.0 % UAR difference from centralized fine-tuning in the worst case. Our conclusions are: (1) In FL distributed computing, smaller pre-trained models fail to fully learn and generalize pre-training knowledge during fine-tuning, leading to performance degradation. Additionally, under the same conditions, small-scale pre-trained models require more task-specific data to achieve optimal performance. However, this is hindered by the dispersed and non-iid data in FL, with limited quantities. (2) Notably, among all fine-tuning methods, LoRA maintains significant performance even in distributed training. This is due to its low-rank decomposition technique, which restricts the weight matrix updates to a low-rank subspace, where the updates are closely related to the specific directions of the original weights (Hu et al., 2022).

4.5. Attack performance

4.5.1. For different pre-trained models and fine-tuning methods

As highlighted in previous work (Mireshghallah et al., 2020; Narra, Lin, Wang, Balasubramanian, & Annavaram, 2021), most information leakage occurs in the early layers of machine learning models, a finding corroborated by our study (Zhao et al., 2023a). In the federated fine-tuning framework, attack performance is assessed using first-layer weight updates and biases from backpropagation. Table 4 compares attack performance across different fine-tuning methods and pre-trained models, using various datasets, with ACC and UAR metrics. FL inherently risks privacy leakage (Feng et al., 2021). However, fine-tuning pre-trained models in FL allows for practical implementation with large-scale models by updating only a subset of parameters. This reduces transmission load and limits sensitive information exposure. The table shows that gender inference results cluster around 50.0 %, with ACC values not exceeding 55.6 %, indicating a chance prediction level. Notably, Whisper-based models show higher inference susceptibility under

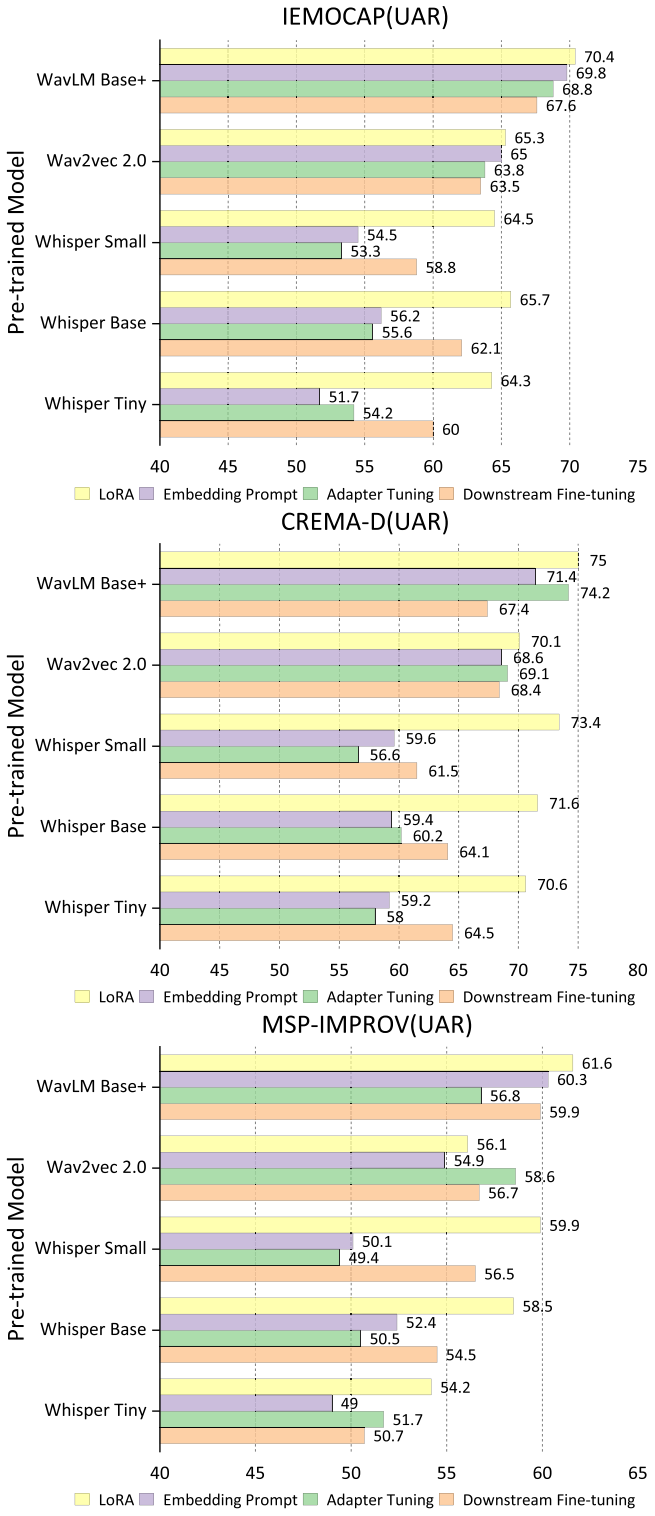

Fig. 6. Performance of downstream fine-tuning with various pre-trained models in FL settings (UAR%).

downstream tuning, whereas Wav2vec 2.0 Base and WavLM Base+ do not exhibit this trend.

4.5.2. Comparison with other works

In comparison, while federated learning fine-tuning slightly underperforms centralized methods in SER, FedLPEFT offers clear privacy advantages by providing a collaborative training mechanism without requiring data sharing or direct access. This builds a strong privacy safeguard. The inclusion of large-scale pre-trained models further enhances FL performance. To illustrate FedLPEFT's advantages, we compared it with existing federated learning approaches under similar conditions. Fig. 7 presents a comprehensive evaluation of SER performance and attribute inference attacks (gender prediction). Feng et al. (2021) implemented a CNN-based federated framework, which showed suboptimal SER performance and attribute inference rates above 80.0 %, exposing significant privacy vulnerabilities. Zhao et al. (2023a) improved privacy strategies to reduce attribute inference to around 52.0 %, achieving modest performance gains overall, but with limited improvement and a downward trend on the CREMA-D dataset. In contrast, our FedLPEFT approach, particularly using WavLM Base+ and LoRA, shows significant improvements in SER performance: approximately 5.0 % better on the IEMOCAP dataset, around 14.0 % and 16.0 % better on two other datasets, demonstrating its effectiveness and robustness. Additionally, we achieve strong privacy protection, with gender prediction UAR not exceeding 52.5 %, ensuring that attackers cannot infer client gender information and providing robust security.

4.6. Trainable parameters

In federated training, factors such as data scale and network bandwidth influence system efficiency, but the size of the pre-trained model and parameter transmission volume are key concerns. Larger models entail more parameters and longer computation times. Utilizing various fine-tuning techniques allows freezing most backbone parameters and updating only a few, thereby reducing communication overhead. To build an effective FedLPEFT, it is crucial to balance the choice of pre-trained models and fine-tuning methods. For reference, Table 5 details the total local trainable parameters for different pre-trained models and fine-tuning methods in federated training. Note that for adapter tuning, embedding prompt, and LoRA methods, the trainable parameters of the downstream model are included, as both need to be shared during training. Additionally, Table 6 compares the total number of trainable parameters across related works, highlighting FedLPEFT with WavLM Base+ and LoRA.

4.7. Parameter sensitivity analysis

In Section 4.1, we systematically compared PEFT methods under centralized and federated settings, using downstream tuning as a baseline. This led to several key conclusions. We followed Feng and Narayanan (2023) for some parameter settings: low-rank dimension of 8, embedding prompt size of 5, and adapter size of 128. We will investigate if these parameters are sensitive and if variations affect results, using the IEMOCAP dataset.

4.7.1. Impact of adapter bottleneck size

This section investigates how adjusting the adapter bottleneck size affects SER performance (UAR). Specifically, we conducted experiments varying the adapter bottleneck size $a \in \{32, 64\}$ while keeping other parameters constant. As shown in Fig. 8, reducing the adapter size improves overall performance for all pre-trained models, with a 2.0 % variance observed. However, the overall trend remains consistent, with WavLM Base+ consistently achieving the best results. When the adapter size is 64, Wav2vec 2.0 Base and WavLM Base+ models perform notably well. In contrast, Whisper series models generally show lower performance; for example, Whisper Base performs slightly better than Whisper Small and Whisper Tiny when the adapter size is 32, but only reaches a UAR of 57.4 %, far behind WavLM Base+ (UAR: 69.1 %). Overall, the adapter size has minimal impact on SER, indicating robustness to hyperparameter variations.

4.7.2. Impact of embedding prompt size

In Section 4.4.2, we observed that embedding prompt and adapter fine-tuning methods performed poorly with Whisper models, significantly below downstream fine-tuning. We tested smaller prompt sizes

Table 3
Performance comparison (UAR%/ACC%) of various fine-tuning strategies with different pre-trained models in centralized vs. FL settings.

Framework	Dataset	Fine-Tune	Pre-trained Model									
			Whisper Tiny		Whisper Base		Whisper Small		Wav2vec 2.0		WavLM +	
			ACC	UAR	ACC	UAR	ACC	UAR	ACC	UAR	ACC	UAR
Centralized	IEMOCAP	Downstream	**65.6**	**67.2**	**66.7**	**68.0**	**65.6**	**67.7**	64.4	66.6	65.7	67.8
		Adapter	62.8	63.9	57.9	58.1	56.5	59.4	59.6	63.8	66.9	69.1
		Prompt	55.2	57.5	57.5	59.9	59.0	59.5	**64.7**	**66.8**	68.4	70.2
		LoRA	63.7	65.8	65.4	67.3	63.1	65.8	63.6	66.3	**69.5**	**70.5**
	CREMA-D	Downstream	**72.0**	**73.2**	74.1	75.8	73.1	76.3	69.9	72.2	77.1	77.4
		Adapter	70.0	71.4	69.6	70.8	62.5	64.8	69.3	71.5	76.8	77.6
		Prompt	62.0	65.0	69.0	71.5	70.4	72.8	73.0	73.5	77.2	78.9
		LoRA	71.9	72.1	**74.7**	**77.0**	**76.2**	**78.4**	**73.4**	**74.8**	**78.9**	**79.3**
	MSP-IMPROV	Downstream	57.8	60.2	**62.8**	**62.7**	**63.4**	**64.2**	57.9	61.4	63.3	62.9
		Adapter	52.2	55.0	51.7	55.8	51.9	51.0	57.5	60.2	62.7	62.5
		Prompt	53.2	53.1	56.9	58.6	57.2	60.0	57.8	61.3	64.8	65.0
		LoRA	**58.5**	**61.1**	61.7	62.3	61.8	61.6	**58.8**	**61.4**	**66.0**	**65.4**
FL	IEMOCAP	Downstream	58.9	60.0	61.7	62.1	56.2	58.8	61.8	63.5	65.5	67.6
		Adapter	53.8	54.2	54.4	55.6	49.8	53.3	61.4	63.8	66.0	68.8
		Prompt	49.7	51.7	55.2	56.2	53.2	54.5	63.6	65.0	67.6	69.8
		LoRA	**63.2**	**64.3**	**64.9**	**65.7**	**63.9**	**64.5**	**63.7**	**65.3**	**69.2**	**70.4**
	CREMA-D	Downstream	66.7	64.5	67.8	64.1	65.0	61.5	69.5	68.4	69.4	67.4
		Adapter	62.0	58.0	67.8	60.2	59.1	56.6	71.3	69.1	76.4	74.2
		Prompt	61.3	59.2	64.0	59.4	61.7	59.6	71.0	68.6	75.5	71.4
		LoRA	**71.7**	**70.6**	**75.8**	**71.6**	**76.8**	**73.4**	**72.1**	**70.1**	**78.2**	**75.0**
	MSP-IMPROV	Downstream	53.0	50.7	55.3	54.5	59.7	56.5	**57.8**	**56.7**	62.8	59.9
		Adapter	55.1	51.7	53.9	50.5	51.8	49.4	60.5	58.6	60.3	56.8
		Prompt	52.0	49.0	56.6	52.4	51.3	50.1	57.0	54.9	63.7	60.3
		LoRA	**58.0**	**54.2**	**61.5**	**58.5**	**61.5**	**59.9**	56.6	56.1	**65.1**	**61.6**

Table 4
Attribute inference attacks analysis (ACC%/UAR%) for different fine-tuning techniques across pre-trained models in FL settings, The ACC and UAR score of the gender prediction task is reported..

D_p	D_ρ	Fine-Tune	Pre-trained Model									
			Whisper Tiny		Whisper Base		Whisper Small		Wav2vec 2.0		WavLM Base+	
			ACC	UAR	ACC	UAR	ACC	UAR	ACC	UAR	ACC	UAR
IEMOCAP		Downstream	**52.2**	**51.0**	**52.1**	**50.9**	**52.5**	**51.3**	50.8	50.7	50.5	49.9
	MSP-IMPROV	Adapter	47.3	49.6	47.0	49.9	47.3	50.0	**52.8**	**50.9**	46.2	46.3
	CREMA-D	Prompt	50.5	50.0	50.4	49.9	50.8	50.2	49.5	48.5	47.0	47.1
		LoRA	46.6	46.5	46.8	46.4	46.4	46.0	51.4	49.4	**51.0**	**50.7**
CREMA-D		Downstream	**54.4**	**52.2**	**54.4**	**52.2**	**54.9**	**52.5**	**55.6**	50.4	52.7	50.0
	IEMOCAP	Adapter	51.7	51.8	51.6	51.8	52.0	52.2	52.5	**51.9**	51.0	50.3
	MSP-IMPROV	Prompt	51.9	51.2	52.0	51.2	52.2	51.5	53.0	50.1	**55.1**	**52.7**
		LoRA	52.2	51.2	52.1	51.1	52.5	51.5	53.2	51.0	54.1	52.5
MSP-IMPROV		Downstream	**54.4**	**53.9**	**54.5**	**53.7**	**55.0**	54.2	51.3	51.2	51.8	51.7
	IEMOCAP	Adapter	52.7	52.3	53.0	52.2	53.2	52.7	51.8	52.3	**52.6**	**53.6**
	CREMA-D	Prompt	50.6	50.8	51.2	51.4	51.1	51.2	**52.3**	**53.0**	52.3	53.3
		LoRA	54.0	53.9	54.0	53.7	54.5	**54.4**	52.1	52.3	51.2	51.3

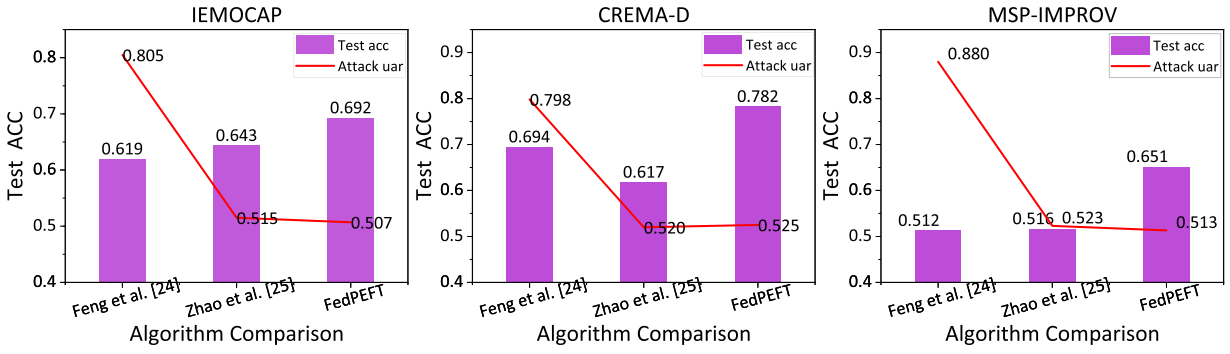

Fig. 7. Comparison of LoRA-based federal parameter-efficient fine-tuning with WavLM Base+ pre-trained models in SER performance and gender inference under FL settings.

Table 5
Trainable parameters under various fine-tuning methods.

Pre-trained Model	Downstream	Adapter	Prompt	LoRA
Whisper Tiny	0.30 M	0.69 M	0.30 M	0.42 M
Whisper Base	0.33 M	1.12 M	0.35 M	0.58 M
Whisper Small	0.40 M	2.77 M	0.44 M	1.13 M
Wave2vec2.0 Base	0.40 M	2.77 M	0.44 M	1.13 M
WavLM Base+	0.40 M	2.77 M	0.44 M	1.13 M

Table 6
Demonstration of trainable parameters for related works.

Comparative Works	Trainable Parameters
Feng et al.(CNN)	0.29 M
Zhao et al.(CNN+BiGRU)	1.42 M
FedLPEFT(Pretrained WavLM Base+LoRA)	1.13 M

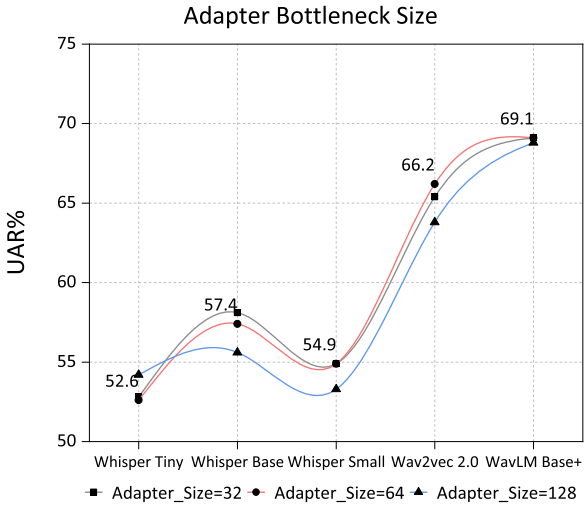

Fig. 8. SER performance with different pre-trained models and adapter bottleneck sizes.

($p \in \{1, 3, 5\}$) and found that reducing prompt size improved SER performance for Whisper models, with minimal impact on Wav2vec 2.0 Base and WavLM Base+ models (Figs. 9). This aligns with conclusions from centralized research (Feng & Narayanan, 2023) and highlights the challenges and careful consideration needed when using embedding prompts in pre-trained model fine-tuning.

4.7.3. Impact of low-rank order

Sections 4.4 and 4.5 underscore LoRA's superior performance among fine-tuning methods, demonstrating its robustness across various pre-trained models, with optimal results achieved on the WavLM Base+ model at rank 8. To assess the impact of low-rank order on SER performance, we conducted experiments with varying low-rank orders $r \in \{8, 16, 32\}$. Results indicate that performance significantly deteriorates for Whisper models when the low-rank order is set to 16 or 32, while Wav2vec 2.0 Base and WavLM Base+ exhibit relative stability (Figs. 10). This suggests that low-rank order is a critical parameter in Whisper model fine-tuning. An increased rank does not necessarily enhance performance; higher rank elevates model complexity and parameter count, potentially complicating optimization. Although a higher rank improves the model's capacity to capture patterns and features, it may also lead to overfitting, impacting generalization. Thus, we recommend careful tuning of low-rank adaptive fine-tuning methods for different pre-trained models and datasets to achieve optimal performance.

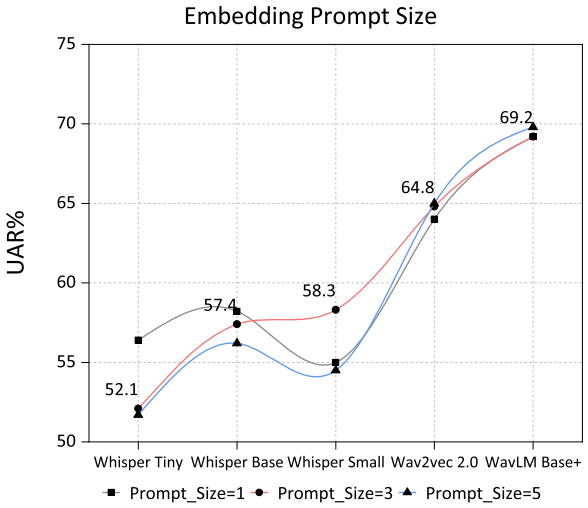

Fig. 9. SER performance across different pre-trained models with varying embedding prompt sizes.

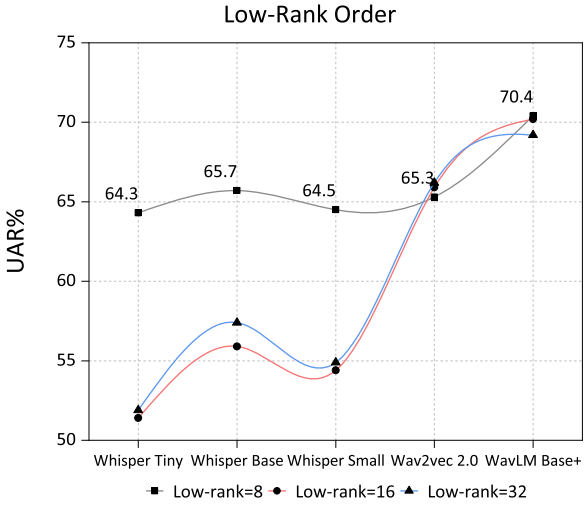

Fig. 10. SER performance across different pre-trained models with varying LoRA sizes.

5. Conclusion

This work develops a PEFT approach within a federated learning framework, which ensures parameter privacy for participants during collaborative training while effectively reducing communication overhead. Specifically, FedLPEFT integrates trainable modules (adapter, embedding prompt, and LoRA) into the feed-forward layers of pre-trained speech models, keeping the backbone network parameters frozen. Fine-tuning these trainable layers improves SER performance, with minimal parameter sharing significantly enhancing privacy and reducing system communication costs. Experimental results demonstrate the substantial performance gains with the WavLM Base+ pre-trained model and the stability brought by LoRA's low-rank decomposition across various scenarios, while effectively mitigating attribute inference attacks. Furthermore, we propose leveraging cloud and edge computing advantages to extend FedLPEFT to integrate smart application scenarios spanning cloud-edge-terminal environments. Future plans include developing more lightweight and secure fine-tuning solutions for larger pre-trained speech models and exploring multimodal knowledge from additional dimensions.

CRediT authorship contribution statement

Haijiao Chen: Methodology, Conceptualization, Formal analysis, Data curation, Software, Writing – original draft, Investigation, Writing – review & editing, Validation; **Huan Zhao:** Funding acquisition, Writing – original draft, Validation, Supervision; **Zixing Zhang:** Investigation, Conceptualization, Writing – original draft, Supervision; **Keqin Li:** Resources, Writing – original draft, Writing – review & editing, Supervision.

Data availability

All data sets used in this paper are publicly accessible.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62076092 and 61772188; and in part by the National Key Research and Development Program of China under Grant 2020YFB1713400.

References

Agrawal, N., Shahin Shamsabadi, A., Kusner, M. J., & Gascón, A. (2019). Quotient: Two-party secure neural network training and prediction. In *Proceedings of the ACM SIGSAC conference on computer and communications security* (pp. 1231–1247).

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of neural information processing systems* (pp. 12449–12460).

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*, 335–359.

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., & Provost, E. M. (2016). MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, *8*(1), 67–80.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, *5*(4), 377–390.

Chandola, D., Altarawneh, E., Jenkin, M., & Papagelis, M. (2024). SERC-GCN: Speech emotion recognition in conversation using graph convolutional networks. In *Proceedings of IEEE international conference on acoustics, speech and signal processing* (pp. 76–80).

Chen, H., Zhao, H., Zhang, Z., & Li, K. (2024). Discriminative feature learning-based federated lightweight distillation against multiple attacks. *IEEE Internet of Things Journal*, *11*(10), 17663–17677.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X. et al. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505–1518.

Chen, Z., Li, J., Liu, H., Wang, X., Wang, H., & Zheng, Q. (2023a). Learning multi-scale features for speech emotion recognition with connection attention mechanism. *Expert Systems with Applications*, *214*, 118943.

Chen, Z., Lin, M., Wang, Z., Zheng, Q., & Liu, C. (2023b). Spatio-temporal representation learning enhanced speech emotion recognition with multi-head attention mechanisms. *Knowledge-Based Systems*, *281*, 111077.

Collins, L., Hassani, H., Mokhtari, A., & Shakkottai, S. (2022). FedAvg with fine tuning: Local updates lead to representation learning. In *Proceedings of neural information processing systems* (pp. 10572–10586).

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W. et al. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, *5*(3), 220–235.

Dixit, C., & Satapathy, S. M. (2024). Deep CNN with late fusion for real time multimodal emotion recognition. *Expert Systems with Applications*, *240*, 122579.

Feng, T., Hashemi, H., Hebbar, R., Annavaram, M., & Narayanan, S. S. (2021). Attribute inference attack of speech emotion recognition in federated learning settings. arXiv preprint arXiv:2112.13416 .

Feng, T., & Narayanan, S. (2023). PEFT-SER: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models. In *Proceedings of international conference on affective computing and intelligent interaction* (pp. 1–8).

Feng, T., Peri, R., & Narayanan, S. (2022). User-level differential privacy against attribute inference attack of speech emotion recognition on federated learning. In *Proceedings of the international speech communication association* (pp. 5055–5059).

Gao, Z.-F., Zhou, K., Liu, P., Zhao, W. X., & Wen, J.-R. (2023). Small pre-trained language models can be fine-tuned as large models via over-parameterization. In *Proceedings of the association for computational linguistics* (pp. 3819–3834).

Geng, J., Mou, Y., Li, F., Li, Q., Beyan, O., Decker, S., & Rong, C. (2021). Towards general deep leakage in federated learning. arXiv preprint arXiv:2110.09074.

Gong, M., Zhang, Y., Gao, Y., Qin, A. K., Wu, Y., Wang, S., & Zhang, Y. (2024). A multi-modal vertical federated learning framework based on homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, *19*, 1826–1839.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *Proceedings of international conference on machine learning* (pp. 2790–2799).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *Proceedings of international conference on learning representations*.

Jaiswal, M., & Provost, E. M. (2020). Privacy enhanced multimodal neural representations for emotion recognition. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 7985–7993).

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., & Lim, S.-N. (2022). Visual prompt tuning. In *Proceedings of European conference on computer vision* (pp. 709–727).

Khan, M., Gueaieb, W., El Saddik, A., & Kwon, S. (2024). MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications*, *245*, 122946.

Kim, H., & Hong, T. (2024). Enhancing emotion recognition using multimodal fusion of physiological, environmental, personal data. *Expert Systems with Applications*, *249*, 123723.

Lashkarashvili, N., Wu, W., Sun, G., & Woodland, P. C. (2024). Parameter efficient finetuning for speech emotion recognition and domain adaptation. In *Proceedings of IEEE international conference on acoustics, speech and signal processing* (pp. 10986–10990).

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. (2021). Survey of deep representation learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, *14*(2), 1634–1654.

Li, D., Liu, J., Yang, Z., Sun, L., & Wang, Z. (2021). Speech emotion recognition using recurrent neural networks with directional self-attention. *Expert Systems with Applications*, *173*, 114683.

Li, T., & Hou, J. (2023). Utilizing self-supervised learning features and adapter fine-tuning for enhancing speech emotion recognition. In *Proceedings of international conference on machine learning, big data and business intelligence* (pp. 79–84).

Li, Y., Mehrish, A., Bhardwaj, R., Majumder, N., Cheng, B., Zhao, S., Zadeh, A., Mihalcea, R., & Poria, S. (2023). Evaluating parameter-efficient transfer learning approaches on sure benchmark for speech understanding. In *Proceedings of IEEE international conference on acoustics, speech and signal processing* (pp. 1–5).

López-Gil, J.-M., & Garay-Vitoria, N. (2024). Assessing the effectiveness of ensembles in speech emotion recognition: Performance analysis under challenging scenarios. *Expert Systems with Applications*, *243*, 122905.

Malaviya, S., Shukla, M., & Lodha, S. (2023). Reducing communication overhead in federated learning for pre-trained language models using parameter-efficient finetuning. In *Proceedings of conference on lifelong learning agents (coLLAs)* (pp. 456–469).

Mireshghallah, F., Taram, M., Ramrakhyani, P., Jalali, A., Tullsen, D., & Esmaeilzadeh, H. (2020). Shredder: Learning noise distributions to protect inference privacy. In *Proceedings of the twenty-fifth international conference on architectural support for programming languages and operating systems* (pp. 3–18).

Naderi, N., & Nasersharif, B. (2023). Cross corpus speech emotion recognition using transfer learning and attention-based fusion of wav2vec2 and prosody features. *Knowledge-Based Systems*, *277*, 110814.

Narra, K. G., Lin, Z., Wang, Y., Balasubramanian, K., & Annavaram, M. (2021). Origami inference: Private inference using hardware enclaves. In *Proceedings of IEEE international conference on cloud computing* (pp. 78–84).

Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *Proceedings of IEEE symposium on security and privacy* (pp. 739–753).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of international conference on machine learning* (pp. 28492–28518).

Ren, Z., Baird, A., Han, J., Zhang, Z., & Schuller, B. (2020). Generating and protecting against adversarial attacks for deep speech-based emotion recognition models. In *Proceedings of IEEE international conference on acoustics, speech and signal processing* (pp. 7184–7188).

Sun, Y., Li, Z., Li, Y., & Ding, B. (2024). Improving loRA in privacy-preserving federated learning. In *Proceedings of the twelfth international conference on learning representations (ICLR)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of neural information processing systems* (pp. 5998–6008).

Wang, X., Wang, M., Qi, W., Su, W., Wang, X., & Zhou, H. (2021). A novel end-to-end speech emotion recognition network with stacked transformer layers. In *Proceedings of IEEE international conference on acoustics, speech and signal processing* (pp. 6289–6293).

Yuan, L., Huang, G., Li, F., Yuan, X., Pun, C.-M., & Zhong, G. (2023). RBA-GCN: Relational bilevel aggregation graph convolutional network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 2325–2337.

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. *Knowledge-Based Systems*, *216*, 106775.

Zhang, Z., Peng, L., Pang, T., Han, J., Zhao, H., & Schuller, B. W. (2024). Refashioning emotion recognition modelling: The advent of generalised large models. *IEEE Transactions on Computational Social Systems*, .

Zhang, Z., Yang, Y., Dai, Y., Wang, Q., Yu, Y., Qu, L., & Xu, Z. (2023). FedPETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In *Proceedings of annual meeting of the association of computational linguistics (ACL)* (pp. 9963–9977).

Zhao, H., Chen, H., Xiao, Y., & Zhang, Z. (2023a). Privacy-enhanced federated learning against attribute inference attack for speech emotion recognition. In *Proceedings of IEEE international conference on acoustics, speech and signal processing* (pp. 1–5).

Zhao, H., Du, W., Li, F., Li, P., & Liu, G. (2023b). FedPrompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. In *Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1–5).

Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. In *Proceedings of neural information processing systems* (pp. 14747–14756).