Semi-FedSER: Semi-supervised Learning for Speech Emotion Recognition On Federated Learning using Multiview Pseudo-Labeling

Tiantian Feng[1], Shrikanth Narayanan[1]

[1]Signal Analysis and Interpretation Lab (SAIL), University of Southern California

`tiantiaf@usc.edu, shri@ee.usc.edu`

Abstract

Speech **E**motion **R**ecognition (**SER**) application is frequently associated with privacy concerns as it often acquires and transmits speech data at the client-side to remote cloud platforms for further processing. These collected speech data reveal not only raw speech content but also the speaker's identity, demographic traits, and health status. Federated learning (FL) is a distributed machine learning algorithm that coordinates clients to train a model collaboratively without sharing local data. This algorithm shows enormous potential for SER applications as sharing raw speech or speech features from a user's device is vulnerable to privacy attacks. However, one major challenge in the FL is limited high-quality labeled data samples. In this work, we propose a Semi-supervised learning framework, Semi-FedSER, that utilizes both the labeled and unlabeled data samples to address the challenge in limited labeled data samples in FL. We show that our Semi-FedSER can generate desired SER performance even when the local label rate $l = 20\%$ using two SER benchmark datasets: IEMOCAP and MSP-Improv.

Index Terms: Speech Emotion Recognition, Semi-supervised Learning, Federated Learning, Pseudo-labeling

1. Introduction

Speech emotion recognition (SER) has gained enormous amount of interests in many applications such as smart virtual assistants [1], medical diagnoses [2, 3], and education [4]. SER is the task that aims to identify emotional states conveyed in vocal expressions automatically. The SER system typically has three parts: data acquisition, data transfer, and classification [5]. However, privacy is a major challenge that researchers or service providers often need to face before deploying the SER system. The speech signals are typically sensitive and can reveal a significant amount of information about an individual. These information includes speech content and speaker's identity, traits (e.g., age, gender), and states (e.g., health status), most of which are deemed sensitive from an application point of view.

Federated learning (FL) is a popular privacy-preserving distributed learning approach that allows clients to train a model collaboratively without sharing their local data [6]. In an FL setting, a central server aggregates model updates from multiple clients during the training process. Each client generates such model updates by locally training a model on the private data available at the client. Furthermore, this machine learning approach reduces information leaks compared to classical centralized machine learning frameworks since personal data does not leave the client. Therefore, this distributed learning paradigm can be a natural choice for developing real-world multiuser SER applications as sharing raw speech or speech features from users' devices is vulnerable to privacy attacks.

Most current works in FL have focused on fully supervised settings where all the input data are with labels. However, high-quality labeled data samples do not often exist in real-life settings, and most data samples are indeed unlabeled. In order to address the limited labeled data samples in the FL setting, prior works have considered using semi-supervised learning (SSL), which utilizes unlabeled samples in addition to a small portion of labeled examples to obtain desired model performance [7, 8]. SemiFed is an SSL framework that integrates supervised and unsupervised training during FL. This approach also generates pseudo-labels by ranking data samples using peer models. In addition, FedMatch [9] uses an inter-device consistency loss to enforce consistency between the pseudo-labeling predictions made across multiple devices. However, we argue that these methods are unrealistic in practice for two reasons: first, SemiFed significantly increases the communication costs occurring at each global training round, where the server needs to transfer additional models to each local client; additionally, sharing the local model to peers may cause potential breaches since the prior work has shown that data reconstruction is possible through accessing model updates [10].

In this work, we propose an SSL framework, Semi-FedSER, to address the challenge in limited labeled data samples for SER application in the FL setting. Semi-FedSER performs the model training, utilizing both labeled and unlabeled data samples at the local client following FixMatch framework [11]. Furthermore, Semi-FedSER addresses the pseudo-labeling issue mentioned earlier using the idea of multiview pseudo-labeling, and we adopt an efficient yet effective data augmentation technique called: Stochastic Feature Augmentation (SFA) [12]. In addition, we propose to utilize the past pseudo label information to improve the quality of the pseudo labels generated. To the best of our knowledge, the only similar work that exists in the literature is [8] for SER application. We show that our Semi-FedSER can generate desired SER performance even when the local label rate $l = 20\%$ and data distribution is completely non-iid using IEMOCAP [13], and MSP-Improv [14] datasets for the experiments.

2. SER Data Sets

In this work, we use two data sets for developing SER models. Table 1 shows the label distribution of utterances in these corpora used in this work.

2.1. IEMOCAP

The IEMOCAP database [13] is a multi-modal corpus collected from ten subjects (five male and five female) who target expressing categorical emotions. The data set contains measurements of motion, audio, and video of acted human interactions. The whole data set contains 10,039 utterances performed in the improvised and scripted conditions. Speakers use the utterance

Table 1: *Statistics of emotion labels in two SER data sets.*

	Neutral	**Happy**	**Sad**	**Angry**	**All**
IEMOCAP	1099	947	608	289	2943
MSP-Improv	2072	1184	739	585	4580

Table 2: *Notation used in this paper.*

t	Global training epoch in FL.
k	Client index.
θ^t	Global SER model at tth global training epoch.
θ_k^t	Local SER model updates of kth client at tth global training epoch.
$\mathbf{D_k^l}$	Labeled speech data of the kth client.
$\mathbf{D_k^P}$	Speech data with hard pseudo labels of the kth client.
$\mathbf{D_k^u}$	Unlabeled speech data of the kth client.
$\mathbf{x^l}$	Labeled speech data.
$\mathbf{x^u}$	Unlabeled speech data.
$\mathbf{x^P}$	Unlabeled speech data with pseudo labels.
$\phi(\cdot)$	Weakly-augmentation function.
$\mathbf{\Phi}(\cdot)$	Strongly-augmentation function.
$\mathbf{y}$	Emotion label.

from a fixed script in the scripted condition, while the utterance in the improvised condition is spontaneous. In this work, we focus on the improvised sessions following the suggestions from [15] and our previous works. Due to the data imbalance issue in the IEMOCAP corpus, previous works use the four most frequently occurring emotion labels (neutral, sad, happiness, and anger) for training the SER model [15]. In addition to this, we pick these four emotion classes because all other corpora contain these labels.

2.2. MSP-Improv

The MSP-Improv [14] corpus was used to study naturalistic emotions captured from improvised scenarios. The corpus includes audio and video data of utterances spoken in natural condition (2,785 utterances), target condition (652 target utterances in an improvised scenario), improvised condition (4,381 utterances from the remainder of the improvised scenarios), and read speech condition (620 utterances). The data is collected from 12 participants (six male and six female). Like the IEMOCAP data set, we only use the data from the improvised conditions.

3. Method

This section describes the details of the proposed semi-supervised learning framework. To improve readability, we summarize the notations adopted in this paper in Table 2.

3.1. Model Overview

In this work, we follow the model framework in FixMatch [11]. We perform two types of augmentations as part of FixMatch: strong and weak, denoted by $\Phi(\cdot)$ and $\phi(\cdot)$, respectively. We generate both weak and strong augmentations through noise addition. Specifically, we adopt the Stochastic Feature Augmentation (SFA) [12] where given a speech feature $\mathbf{x}$, the weak augmentation is $\phi(\mathbf{x}) = \mathbf{x} \odot \alpha + \mathbf{r}, \quad$ where $\alpha \sim \mathcal{N}(1, \sigma_1), \mathbf{r} \sim \mathcal{N}(0, \sigma_2)$. We choose this method as this efficient yet straightforward feature augmentation technique has shown promising results in model generalization tasks [12]. We choose a larger σ_1 in $\Phi(\cdot)$ than in $\phi(\cdot)$.

We formulate the learning objectives to include labeled and unlabeled data information. Specifically, our learning function has a supervised loss term l_s and an unsupervised loss term l_u. The supervised loss on labeled data is the standard cross-entropy loss, and given the kth client and its local model with parameter θ^k, we aim to minimize the following l_s with the weakly-augmented labeled data $\phi(x^l)$:

$$\min \mathcal{L}_{ce}(f(\theta^\mathbf{k}; \phi(\mathbf{x^l})), y) \qquad (1)$$

When modeling the unlabeled data, our method computes a pseudo label for unlabeled data, which is then used in a standard cross-entropy loss. We obtain the pseudo label y' using the unlabeled weakly-augmented data. We explain our pseudo labeling process in the following subsection. We define the cross-entropy loss against the model's output using the strongly-augmented version of unlabeled data $\mathbf{x^P}$:

$$\min \mathcal{L}_{ce}(f(\theta^\mathbf{k}; \mathbf{\Phi}(\mathbf{x^P})), y') \qquad (2)$$

Our final goal is to minimize the combined loss below:

$$\min l_s + l_u \qquad (3)$$

3.2. Pseudo Labeling

Our Semi-FedSER framework adopts the idea of pseudo labeling proposed in [7]. At each iteration, the model uses predictions from the previous iteration as target classes for unlabeled data samples as if they were true labels. The process goes through a fixed number of iterations until all unlabeled data are labeled. However, the generated pseudo labels can be inaccurate and negatively impact the training results. In order to improve the quality of the generated pseudo labels, [7] proposed to send the aggregated model θ^t and K other local models θ_k^t to each client. Then for each local data sample, the local client produces $K + 1$ one-hot labels. This approach generates the final pseudo label only if a certain amount of models agrees with the sample label. However, we argue that this method is highly inefficient and unsafe in practice.

Multiview pseudo-labeling (MvPL) Unlike the pseudo-labeling process in [7], our framework utilizes the idea of the multiview pseudo-labeling (MvPL) approach proposed in [16]. The MvPL takes in multiple complementary views of a single unlabeled data point and uses a shared model to generate predictions. Similarly, we decide the final pseudo label by aggregating predictions from all the views. Here, we generate the weakly-augmented views of a speech feature sample using the SFA approach [12]. Specifically, in our setting, the local client generates m weakly-augmented views using SFA for unlabeled data $\mathbf{x^u}$. Suppose there are m predicted class distributions $\mathbf{q_1}, \mathbf{q_2}, ..., \mathbf{q_m}$, and we can obtain the averaged predicted class distributions as $\bar{\mathbf{q}}$ from all weakly-augmented views. We finally assign the pseudo label y' to $\mathbf{x^u}$ if $\max(\bar{\mathbf{q}})$ is above the confidence threshold τ. To further increase the confidence of a prediction, we apply the temperature scaling with a constant scalar temperature T to soften the logits output from the model. Specifically, the probability output for jth class q_j is:

$$q_j = \frac{e^{z_j/T}}{\sum_{i=1}^{C} e^{z_i/T}} \qquad (4)$$

Compared to the pseudo-labeling process in [7], our method is more efficient and does not require sharing local models.

Pseudo-labeling uncertainty To further increase the quality of the pseudo labels, we propose to use the idea of uncertainty-aware pseudo label selection process (UPS) [17]. This pseudo-

label selection process utilizes both the confidence and uncertainty of a network prediction. We obtain the uncertainty measure $u(\mathbf{q})$ by calculating the standard deviation of m predicted class distributions. We compute the loss for an unlabeled strongly-augmented data $\mathbf{\Phi}(\mathbf{x^u})$ as:

$$\mathbb{1}(\bar{\mathbf{q}} \geq \tau) \cdot \mathbb{1}(u(\mathbf{q}) \leq k) \cdot \mathcal{L}_{ce}(f(\theta^\mathbf{k}; \mathbf{\Phi}(\mathbf{x^u})), \mathbf{y'}) \quad (5)$$

Finally, we add data of the pseudo labels to the pseudo-label data and delete them from the unlabeled data. The entire pseudo labeling algorithm is presented in the algorithm 1.

Algorithm 1 MvPL with pseudo-labeling consistency

1: **Input:** $\mathbf{D_k^p}, \mathbf{D_k^u}, \theta_\mathbf{k}^{\mathbf{t'}}, m, \tau, k$
2: **for** each client $k \in S$ in parallel **do**
3:     **for** $x^u \in \mathbf{D_k^u}$ **do**
4:         $\mathbf{q} \leftarrow []$
5:         **for** $i = 1, 2, ..., m$ **do**
6:             $\phi_i(\mathbf{x^u}) = \mathbf{x^u} \odot \alpha + \mathbf{r}$ where $\alpha \sim \mathcal{N}(1, \sigma_1)$,
7:             $\mathbf{r} \sim \mathcal{N}(0, \sigma_2)$
8:             append $f(\phi_i(\mathbf{x}^u); \theta^t)$ to $\mathbf{q}$
9:         $\bar{q} \leftarrow \text{average}(\mathbf{q})$
10:        $u(\mathbf{q}) \leftarrow \text{std}(\mathbf{q})$
11:        $y' \leftarrow \arg\max(\bar{q})$
12:        **if** $\bar{q} \geq \tau$ and $u(\mathbf{q}) \leq k$ **then**
13:            $\mathbf{D_k^p} \cup \mathbf{x}^u; \mathbf{D_k^u} \setminus \mathbf{x}^u$

3.3. SCAFFOLD

We often encounter non-IID data at different clients in a typical FL setting. Thus, gradient drifting is a major challenge when training the FL algorithm in a non-IID data setup. To mitigate the impact of the gradient drifting, we propose to implement SCAFFOLD [18]. SCAFFOLD is a robust algorithm that uses control variates $\mathbf{c}$ (variance reduction) to correct for the 'client-drift' in its local updates. The details of the algorithm proof are in [18]. Finally, the algorithm 2 shows the overall training procedure for our proposed Semi-FedSER pipeline.

4. Experiments

In this section, we describe our experimental setup and training details. The implementation of this work is at https://github.com/usc-sail/fed-ser-semi.

4.1. Data Preprocessing

To investigate the effectiveness of the proposed semi-supervised learning framework, we train our SER models on various speech representations. We first decide to evaluate our proposed pipeline using the knowledge-based speech feature set, Emo-Base feature set, computed from the OpenSMILE toolkit [19]. In addition to the knowledge-based speech feature set, we propose to test our framework on SUPERB (Speech Processing Universal PERformance Benchmark) [20], which is designed to provide a standard and comprehensive testbed for pre-trained models on various downstream speech tasks. We compute the deep speech representations from the pre-trained models that are available in SUPERB including APC [21], TERA [22], and DeCoAR 2.0 [23] and DistilHuBERT [24]. We choose these three representations as they show better SER performance in our prior work [25]. Finally, we calculate the global average of the last layer's hidden state as the final feature from the pre-trained model's output. Using the last hidden state is suggested

---

Algorithm 2 Semi-FedSER

1: **Initialize:** $\theta^\mathbf{0}, \mathbf{c^0}$
2: **Server executes:**
3: **for** Each round $t = 0, ..., T - 1$ **do**
4:     Sample clients $\mathcal{S} \in \{1, 2, ..., K\}$
5:     **for** Each client $k \in \mathcal{S}$ in parallel **do**
6:         $\theta_\mathbf{k}^\mathbf{t} \leftarrow \theta^\mathbf{t}$
7:         $\theta_\mathbf{k}^\mathbf{t}, \mathbf{c_k}, \nabla\mathbf{c_k} \leftarrow \text{ClientLocalTraining}(\theta_\mathbf{k}^\mathbf{t}, \mathbf{c^t}, \mathbf{c_k})$
8:     $\theta^{\mathbf{t+1}} \leftarrow \frac{1}{|\mathcal{S}|} \sum_{k \in \mathcal{S}} \theta_\mathbf{k}^\mathbf{t}$
9:     $\mathbf{c^{t+1}} \leftarrow \mathbf{c^t} + \frac{1}{|\mathcal{S}|} \sum_{k \in \mathcal{S}} \nabla\mathbf{c_k}$
10: **function** CLIENTLOCALTRAINING($\theta, \mathbf{c}, \mathbf{c'}$)
11:     $\theta_\mathbf{old} \leftarrow \theta$
12:     $\mathbf{c_{old}} \leftarrow \mathbf{c'}$
13:     **for** Local epoch e from 0 to $E - 1$ **do**
14:         **for** Iteration i from 0 to $I - 1$ **do**
15:             Sample mini-batch l from $\mathbf{D_k^l}$
16:             Sample mini-batch $\mathbf{p}$ from $\mathbf{D_k^P}$
17:             $\theta \leftarrow \theta - \eta\nabla_\theta(\mathcal{L}_{ce}(f(\theta; \phi(\mathbf{x^l})), y) +$
18:                 $\mathcal{L}_{ce}(f(\theta; \mathbf{\Phi}(\mathbf{x^P})), y')$
19:             $\theta \leftarrow \theta - \eta(\mathbf{c} - \mathbf{c'})$
20:     **Execute** Algorithm 1
21:     $\mathbf{c'} \leftarrow \mathbf{c'} - \mathbf{c} + \frac{1}{EI\eta}(\theta_\mathbf{old} - \theta)$
22:     **end** $\nabla\mathbf{c} = \mathbf{c'} - \mathbf{c_{old}}$
23:     **return** $\theta, \mathbf{c'}, \nabla\mathbf{c}$

---

in prior works for downstream tasks [21, 26, 23, 27]. Our feature sizes are 988 in Emo-Base; 512 in APC; 768 in Tera, DistilHuBERT, and DeCoAR 2.0. We also apply z-normalization to the speech features within each client.

4.2. Non-IID Data setup

For simulating non-IID distributions over the clients in this experiment, we group the data by each speaker in the IEMOCAP, and MSP-Improv data set and split each speaker's data into 4 different shards. As a result, we produce pathological non-IID data for each partition by taking only 3 unique emotions from each speaker. We regard each partition as a client in federated learning. Consequently, each client in federated learning automatically gets 3 different emotions from one speaker. About 80% of the speakers are for training, and the rest are for the test set. 20% of the data samples in each training client are used as validation, and the rest are training samples. We test our proposed semi-supervised learning pipeline at the labeling rate (training samples) of 20% and 40%. We repeat the experiments 5 times with test folds of different speakers in each data set, and we report the average results of the 5-fold experiments.

4.3. Model and training Details

We use the multilayer perceptron (MLP) as the SER model architecture in this work. The model consists of 2 dense layers with layer sizes of $\{256, 128\}$. We choose ReLU as the activation function and the dropout rate as 0.2. We implement the SCAFFOLD algorithm in training the SER model. We conduct the experiments on a computer with two NVIDIA GeForce RTX 5000 GPUs to report the performance. We control only 10% of the clients participating in each global round. 80% of the data at a client is for local training, and the rest 20% is for validation. We set the local training batch size as 16. Expressly, we set the learning rate as 0.0001 and the local training epoch as 1 in the FL algorithm. The total global training epochs are 500.

Table 3: *Prediction results of the SER model trained in different scenarios. The unweighted average recall (UAR) scores of the SER task on individual data set are reported. The feature sets include Emo-Base obtained from the OpenSMILE toolkit and APC, Tera, and DeCoAR 2.0 computed using pre-trained models. L represents the label rate at the local client.*

| Data set | Feature | Fully Supervised | | | L=20% | | L=40% | |
		Centralized	FedAvg	SCAFFOLD	Supervised	Semi-FedSER	Supervised	Semi-FedSER
IEMOCAP	**Emo-Base**	**61.83%**	58.01%	61.34%	58.49%	**60.13%**	**59.60%**	59.51%
	APC	**66.70%**	62.07%	66.21%	61.21%	**63.11%**	64.98%	**65.48%**
	DeCoAR 2.0	**65.82%**	62.22%	65.09%	58.63%	**61.14%**	62.00%	**63.77%**
	Tera	**65.41%**	60.17%	64.62%	56.85%	**59.11%**	63.12%	**63.69%**
	DistilHuBERT	**68.17%**	62.80%	67.81%	61.41%	**62.19%**	62.65%	**64.10%**
MSP-Improv	**Emo-Base**	**48.09%**	45.85%	47.73%	45.56%	**46.62%**	46.20%	**47.46%**
	APC	**53.57%**	48.33%	52.91%	48.55%	**50.33%**	50.71%	**50.82%**
	DeCoAR 2.0	**53.99%**	49.38%	53.37%	47.70%	**49.16%**	51.73%	**52.05%**
	Tera	**53.82%**	50.38%	53.81%	49.26%	**50.11%**	50.91%	**51.66%**
	DistilHuBERT	**55.65%**	51.22%	55.37%	50.51%	**50.74%**	52.01%	**53.72%**

We set σ_2 as 0.1 when generating weakly-augment samples and strongly-augment samples. We choose σ_1 as 0.1 and 0.25 in generating weakly-augment samples and strongly-augment samples, respectively. We use a confidence score $\tau = 0.5$ in the MvPL algorithm at the beginning of the training and linearly increase it to 0.9 at the 300th global training epoch. We apply a temperature value of 2 to soften the prediction output from weakly-augment samples during the MvPL. We empirically use the number of augmentations m and uncertainty score k as 10 and 0.005, respectively. To balance the pseudo label distribution, we add only one data sample with pseudo labels for each class in each training epoch.

5. Results

5.1. Compared baselines

We compare our semi-supervised learning framework with the following baselines:

Supervised - Federated Here, we refer to supervised baseline as using only labeled data points at each local client with an assigned labeling rate. We train the supervised Federated baseline using the SCAFFOLD algorithm.

Fully Supervised - Federated We train a fully supervised Federated baseline (use all the original labels of the training data) to measure the ideal accuracy we would obtain under federated learning settings. We train the fully supervised Federated baseline using both FedAvg and SCAFFOLD.

Fully Supervised - Centralized Finally, we train a fully supervised centralized baseline to measure the upper bound SER model performance in different data sets.

5.2. SER performance

Baselines The emotion prediction results of FL training on different data sets are shown in Table 3. We report the SER prediction results in unweighted average recall (UAR) scores. We observe that our fully-supervised federated SER performance using the SCAFFOLD algorithm yields similar performance to prior works in [28, 29]. Moreover, we can identify that the performance of the fully supervised SER model trained in a centralized setup is slightly higher than the fully supervised SER model trained using the SCAFFOLD algorithm. Meanwhile, we find that the SCAFFOLD algorithm performs consistently better than the FedAvg algorithm in training the fully super-

vised SER model. These results suggest that the SCAFFOLD algorithm can significantly increase the model performance in the federated learning setup.

Semi-FedSER When the label rate of the local client is at 20%, we can observe that the SER performance of the federated supervised model drops 5-10% compared to the fully federated supervised model in different data sets and using different speech features. However, we can observe that SER models trained using our proposed Semi-FedSER framework generates similar UAR scores to fully federated supervised models. For example, the UAR score of the fully supervised model using FL is 66.70% in the IEMOCAP data set using the APC speech feature. In contrast, the SER model trained using our proposed Semi-FedSER framework on the same speech feature produces a similar UAR score of 63.11%. We can further observe that the performance of the Semi-FedSER framework is at the same level as the fully supervised FL model across all testing scenarios when the label rate is 40%. These comparisons demonstrate that the proposed Semi-FedSER can achieve desirable performance with few on-device labeled samples.

6. Conclusions

In this work, we propose a novel Semi-FedSER framework to address the challenge in limited labeled data samples for speech emotion recognition application in FL setting. Semi-FedSER leverages both labeled and unlabeled data samples at the local client combined with pseudo-labeling. The pseudo-labeling approach in Semi-FedSER is based on the multiview pseudo-labeling. We further improve the quality of the pseudo-labeling by incorporating the pseudo-labeling consistency measurement. We also implement the SCAFFOLD algorithm to address the non-IID data distribution issue in the FL setting. Our results show that the proposed Semi-FedSER framework effectively generates desired SER predictions even when the local label rate $l = 20\%$. In the future, we plan to extend our current work to multi-modal data scenarios where not every modality is presented with labels. We also aim to design an automatic scheme for deciding the threshold τ in generating pseudo labels.

7. Acknowledgments

8. References

[1] M.-C. Lee, S.-Y. Chiang, S.-C. Yeh, and T.-F. Wen, "Study on emotion recognition and companion chatbot using deep neural network," *Multimedia Tools and Applications*, vol. 79, no. 27, pp. 19 629–19 657, 2020.

[2] S. Ramakrishnan and I. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, 2013.

[3] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, and S. Narayanan, "Signal processing and machine learning for mental health research and clinical applications," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 189–196, September 2017.

[4] W. Li, Y. Zhang, and Y. Fu, "Speech emotion recognition in e-learning system based on affective computing," in *Third International Conference on Natural Computation (ICNC 2007)*, vol. 5. IEEE, 2007, pp. 809–813.

[5] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[6] 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.

[7] H. Lin, J. Lou, L. Xiong, and C. Shahabi, "Semifed: Semi-supervised federated learning with consistency and pseudo-labeling," *arXiv preprint arXiv:2108.09412*, 2021.

[8] V. Tsouvalas, T. Ozcelebi, and N. Meratnia, "Privacy-preserving speech emotion recognition through semi-supervised federated learning," *arXiv preprint arXiv:2202.02611*, 2022.

[9] W. Jeong, J. Yoon, E. Yang, and S. J. Hwang, "Federated semi-supervised learning with inter-client consistency & disjoint learning," *arXiv preprint arXiv:2006.12097*, 2020.

[10] L. Zhu and S. Han, "Deep leakage from gradients," in *Federated learning*. Springer, 2020, pp. 17–31.

[11] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, 2020.

[12] P. Li, D. Li, W. Li, S. Gong, Y. Fu, and T. M. Hospedales, "A simple feature augmentation for domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8886–8895.

[13] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[14] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[15] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1771–1775.

[16] B. Xiong, H. Fan, K. Grauman, and C. Feichtenhofer, "Multiview pseudo-labeling for semi-supervised learning from video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7209–7219.

[17] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," *arXiv preprint arXiv:2101.06329*, 2021.

[18] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[20] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[21] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," in *Interspeech*, 2019.

[22] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[23] S. Ling and Y. Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," 2020.

[24] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," *arXiv preprint arXiv:2110.01900*, 2021.

[25] T. Feng, H. Hashemi, R. Hebbar, M. Annavaram, and S. S. Narayanan, "Attribute inference attack of speech emotion recognition in federated learning settings," *arXiv preprint arXiv:2112.13416*, 2021.

[26] A. Liu, Y.-A. Chung, and J. Glass, "Non-autoregressive predictive coding for learning speech representations from local dependencies," *arXiv preprint arXiv:2011.00406*, 2020.

[27] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[28] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *Interspeech*, 2017, pp. 1089–1093.

[29] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 126–131.