



Federated Self-training for Semi-supervised Audio Recognition

VASILEIOS TSOUVALAS, AAQIB SAEED, and TANIR OZCELEBI, Eindhoven University of Technology, The Netherlands

Federated Learning is a distributed machine learning paradigm dealing with decentralized and personal datasets. Since data reside on devices such as smartphones and virtual assistants, labeling is entrusted to the clients or labels are extracted in an automated way. Specifically, in the case of audio data, acquiring semantic annotations can be prohibitively expensive and time-consuming. As a result, an abundance of audio data remains unlabeled and unexploited on users' devices. Most existing federated learning approaches focus on supervised learning without harnessing the unlabeled data. In this work, we study the problem of semi-supervised learning of audio models via self-training in conjunction with federated learning. We propose FedSTAR to exploit large-scale on-device unlabeled data to improve the generalization of audio recognition models. We further demonstrate that self-supervised pre-trained models can accelerate the training of on-device models, significantly improving convergence within fewer training rounds. We conduct experiments on diverse public audio classification datasets and investigate the performance of our models under varying percentages of labeled and unlabeled data. Notably, we show that with as little as 3% labeled data available, FedSTAR on average can improve the recognition rate by 13.28% compared to the fully supervised federated model.

CCS Concepts: • **Computing methodologies** → **Semi-supervised learning settings**; **Neural networks**; • **Human-centered computing** → *Ubiquitous and mobile computing*;

Additional Key Words and Phrases: Federated learning, semi-supervised learning, deep learning, audio classification, sound recognition, self-supervised learning

ACM Reference format:

Vasileios Tsouvalas, Aaqib Saeed, and Tanir Ozcelebi. 2022. Federated Self-training for Semi-supervised Audio Recognition. *ACM Trans. Embed. Comput. Syst.* 21, 6, Article 74 (October 2022), 26 pages.
<https://doi.org/10.1145/3520128>

1 INTRODUCTION

The emergence of smartphones, wearables, and modern **Internet of Things (IoT)** devices results in a massive amount of highly informative data generated continuously from a multitude of embedded sensors and logs of user interactions with various applications. The ubiquity of these contemporary devices and the exponential growth of the data produced on edge provides a unique opportunity to tackle critical problems in various domains, such as healthcare, well-being, manufacturing, and infrastructure monitoring. Notably, the advent of deep learning has enabled

Authors' address: V. Tsouvalas, A. Saeed, and T. Ozcelebi, Eindhoven University of Technology, Eindhoven, The Netherlands; emails: {v.tsouvalas, a.saeed, t.ozcelebi}@tue.nl.



This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2022 Copyright held by the owner/author(s).
1539-9087/2022/10-ART74
<https://doi.org/10.1145/3520128>

us to leverage these raw data directly for learning models while leaving ad hoc (hand-designed) approaches largely redundant. The improved schemes for learning deep networks and the availability of massive labeled datasets have brought tremendous advancements in several areas, including language modeling, audio understanding, object recognition, image synthesis, and more. Traditionally, developing machine learning models or performing analytics in a data center context requires the data from IoT devices to be pooled or aggregated in a centralized repository before processing it further for the desired objective. However, the rapidly increasing size of available data, in combination with the high communication costs and possible bandwidth limitations, render the accumulation of data in a cloud-based server unfeasible [24]. Additionally, such centralized data aggregation schemes could also be restricted by privacy issues and regulations (e.g., General Data Protection Regulation). Due to these factors and the growing computational and storage capabilities of distributed devices, it is appealing to leave the data decentralized and perform operations directly on the device that collects that data through primarily utilizing local resources. The rapidly evolving **Federated Learning (FL)** field is concerned with distributed training of machine learning models on the decentralized data residing on remote devices such as smartphones and wearables. The key idea behind FL is to bring the computation (or code) closer to where the data reside to harness data locality extensively. Specifically, in a federated setting, minimal updates to the models (e.g., parameters of a neural network) are performed entirely on-device and communicated to the central server, which aggregates these updates from all participating devices to produce a unified global model. Unlike the standard way of learning models, the salient differentiating factor is that the data never leaves the user's device, which is an appealing property for privacy-sensitive data. This strategy has been applied on a wide range of tasks in recent years [10, 23, 34, 44]. Nevertheless, a common limitation of existing approaches is that they primarily focus on a supervised learning regime. The implicit assumption that the labeled data is widely available on the device, or it can be easily labeled through user interaction or programmatically, such as for keyword prediction or photo categorization, is in most pragmatic cases unrealistic.

In reality, on-device data is largely unlabeled and constantly expanding in size. It cannot be labeled to the same extent as standard datasets, which are annotated via crowd-sourcing or other means for training deep neural networks. Due to the prohibitive cost of annotation, users have little to no incentives, and notably for various important tasks, the domain knowledge missing to perform the annotation process appropriately leaves most of the data residing on devices to remain unlabeled. This is especially true when considering the utilization of audio data to perform various audio recognition tasks, which have recently attracted increasing interest from researchers. As a result, numerous audio recognition systems have been developed, such as for wildlife monitoring [28, 36] and surveillance [5]. In addition to monitoring applications, highly accurate acoustic models are utilized for keyword spotting for virtual assistants [23], anomaly detection for machine sounds [18], and in the development of health risk diagnosis systems, such as cardiac arrest detection [4]. However, in the majority of such applications, there is no straightforward manner for the annotation process. For instance, suppose that we have a sleep tracker application that assesses a person's risk of obstructive sleep apnea using breathing and snoring sounds during sleep. In this case, the end-users may not be able to evaluate their sleeping sounds sufficiently, and clinicians may need to analyze and annotate the samples. Even in cases where no human expertise is required, like in a music tagging application, the correct labeling of songs requires effort on the user's end. Additionally, there are cases where distributed devices host models with no human-in-the-loop to annotate the audio data, such as surveillance devices, making the labeling process infeasible. Thus, in many realistic scenarios for FL, local audio data will be primarily unlabeled. This leads to a novel FL problem, namely, *semi-supervised federated learning*, where users' devices

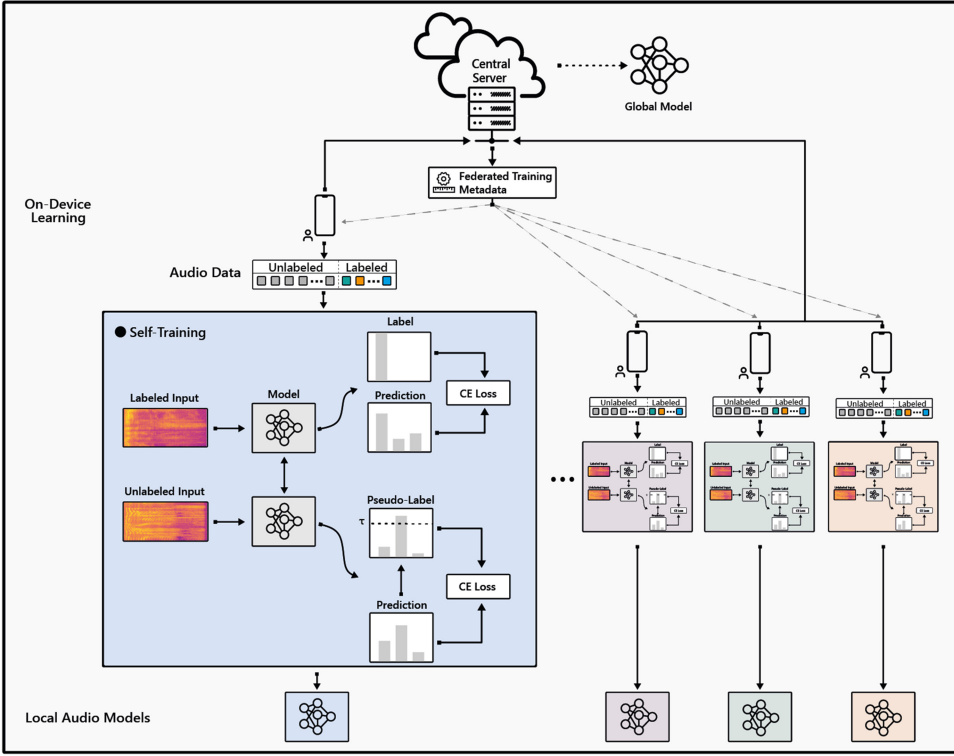


Fig. 1. Illustration of FedSTAR for label-efficient learning of audio recognition models in a federated setting.

collectively hold a massive amount of unlabeled audio samples and only a fraction of labeled audio examples.

Semi-supervised learning techniques have been widely deployed in a centralized learning setting to utilize readily available unlabeled data and could also be applied in federated learning settings. In particular, with semi-supervision of models, available unlabeled data can be exploited during the training phase, improving the overall performance of the resulting model [40]. Pseudo-labeling is a widely applied semi-supervised learning method, which relies on the predictions of a model on unlabeled data, i.e., pseudo labels, to utilize unlabeled data during the learning phase [22]. With no structural requirements from the input modalities and tiny computational overhead, pseudo-labeling is an ideal candidate to be applied in federated learning settings, where device heterogeneity and computational resources vary across devices. To this end, we propose a federated self-training approach, named **FedSTAR (Federated Self-Training for Audio Recognition)**, to unify semi-supervision with federated learning to leverage large-scale unlabeled audio data. With the exploitation of unannotated audio samples that reside on clients' devices, we aim to improve the generalization of federated models on a wide range of audio recognition tasks under a pragmatic scenario, where scarcity of labels poses a significant challenge for learning useful models.

Apart from the labels' deficiency, FL introduces other challenges of the system and statistical heterogeneity [17]. These challenges lead to device hardware and data collection diverseness that can significantly affect the number of devices participating in each federated round as well as the on-device data distribution. Several FL techniques provide flexibility in selecting a fraction of clients in each training round and address the non-i.i.d. nature of client's data

distributions, such as FedAvg [30] and FedProx [25]. The training convergence properties of such distributed optimization methods are discussed in Reference [17], where a clear reduction in the convergence rates is reported. In a centralized setting, self-supervised pre-training can improve the model's convergence and generalization through leveraging pre-training on massive unlabeled datasets [35]. With self-supervised learning, the model is able to learn useful representations from unlabeled data; thus, when used for the downstream task, self-supervised model can significantly improve the training efficiency and predictive performance [35]. To address the issue of slow training convergence in federated settings, we propose the utilization of self-supervised pre-trained models as model initialization for the FL procedure as compared to the naive random initialization of model parameters. Through extensive evaluation, we demonstrate that the convergence rate of our proposed semi-supervised federated algorithm, i.e., FedSTAR, can be greatly improved by using a pre-trained model learned in a self-supervised manner.

To the best of our knowledge FedSTAR is the first FL approach that learns models for audio recognition tasks by utilizing not only labeled but also unlabeled samples on user devices while not being dependent on any data (labeled or unlabeled) on the server side. Just like the labeled samples, the on-device unlabeled samples are utilized locally by self-training based on our proposed pseudo-labeling with dynamic prediction confidence thresholding. As FedSTAR is not altering either the utilized model's architecture or the global model averaging process, the underlined hardware requirements are similar to the chosen federated learning algorithm (e.g., FedAvg), while the on-device storage demand is unaffected, since FedSTAR essentially uses already stored unlabeled data that are left unexploited. In addition, with the utilization of unlabeled data, FedSTAR models are less sensitive to the non-i.i.d. nature of the labeled data across clients (label distribution skew and data sample imbalance across clients). As a result, it performs much better in typical non-i.i.d. data federated settings. Furthermore, solutions in the literature focus on randomly initialized models at the server side. We for the first time employ self-supervised pre-training on the server side using publicly available audio data to further improve the efficiency of training, which means fewer training rounds are needed for convergence.

The main contributions of this work are as follows:

- We study on the practical problem of semi-supervised federated learning for audio recognition tasks to address the lack of labeled data that presents a major challenge for learning on-device models.
- We design a simple yet effective approach based on self-training, called FedSTAR. It exploits large-scale unlabeled distributed data in a federated setting with the help of a novel adaptive confidence thresholding mechanism for effectively generating pseudo-labels.
- We exploit self-supervised models pre-trained on FSD-50K corpus [6] for significantly improving training convergence in federated settings.
- We demonstrate through extensive evaluation that our technique is able to effectively learn generalizable audio models under a variety of federated settings and label availability on diverse public datasets, namely, Speech Commands [41], Ambient Context [33], and VoxForge [29].
- We show that FedSTAR, with as few as 3% labeled data, on average can improve recognition rate by 13.28% across all datasets compared to the fully supervised federated models.

The rest of the article is organized as follows: In Section 2, an overview of the related work is provided. Section 3 presents an overview of related paradigms and methodologies as background information, Section 4 introduces the proposed federated self-training approach for semi-supervised audio recognition. Section 5 presents an evaluation of FedSTAR on publicly available datasets. Section 6 concludes the article and lists future directions for research.

2 RELATED WORK

Federated Learning. FL has been attracting growing attention, thanks to its unique characteristic of collaboratively training machine learning models without actually sharing local data and compromising users' privacy [19]. The most popular and simplistic approach to learning models from decentralized data is the **Federated Averaging (FedAvg)** algorithm [30]. Specifically, FedAvg performs several local **stochastic gradient descent (SGD)** steps on a sampled subset of devices' data in parallel and aggregates the locally learned model parameters on a central server to generate a unified global model through weighted averaging. This strategy has proved to work relatively well for a wide range of tasks in i.i.d. settings [23, 44]. At the same time, the performance can decrease substantially when FedAvg is exposed to non-i.i.d. data distribution [17, 45]. Authors in Reference [45] proposed globally sharing a portion of the dataset to improve FL performance under non-i.i.d. settings. In addition to the challenge introduced by data distribution, communication efficiency is another critical problem in FL. The communication challenges could be alleviated by increasing the number of local SGD steps between sequential communication stages. However, with the increase of SGD steps, the device's model may begin to diverge, and the aggregation of such models can affect the generalization of global models [25]. FedProx was proposed to tackle this issue by adding a loss term to restrict the local models' updates to be closer to the existing global model [25]. Nevertheless, a typical limitation of existing work is the focus on a supervised learning regime with the implicit assumption that the local private data is fully labeled or could be labeled simplistically through labeling functions. However, in the majority of pragmatic scenarios, a straightforward annotation process is non-existent.

Recently, performing on-device federated training of acoustic models has attracted considerable attention [7, 9, 12, 23, 44]. In Reference [23], FL was employed for a keyword spotting task and the development of a wake-word detection system, whereas, References [7, 9] investigated the effect of non-i.i.d. distributions on the same task. In Reference [7], a highly skewed data distribution scenario was considered, where a large set of speakers used their devices to record a set of sentences. To address the challenges introduced due to the non-i.i.d. distribution of data, a word-error-rate model aggregation strategy was developed. In addition, a training scheme with a centralized model, pre-trained on a small portion of the dataset, was also examined. Furthermore, Reference [9] considered a scenario where devices might hold unlabeled audio samples and used a semi-supervised federated scheme based on a teacher-student architecture to exploit unlabeled audio data. However, the teacher model relied on additional high-quality labeled data for training in a centralized setting. Likewise, Reference [12] introduced a framework for privacy-preserving training of user authentication models with FL using labeled audio data. Nonetheless, all prior approaches consider only semantically annotated audio examples or require supplementary labeled data on the server-side to utilize the available unlabeled audio data that reside on devices. To address these problems, we propose a self-training approach to exploit unlabeled audio samples residing on clients' devices. In addition, as servers often possess the computational resources to efficiently pre-train a model on a massive unlabeled dataset, we employ self-supervision to develop a model that can be used as a highly effective starting point for federated training instead of using randomly initialized weights.

Semi-Supervised Learning. In semi-supervised learning (SSL), we are provided with a dataset containing both labeled and unlabeled examples, where the labeled fraction is generally tiny compared to the unlabeled one and the curation of strong labels for the unlabeled dataset is impractical due to time constraints, cost, and privacy-related issues [46]. While there is a wide range of SSL methods and approaches that have been developed in the area of deep learning, we will mainly focus on the self-training or pseudo-labeling approach [22]. Self-training uses the prediction on unlabeled data to supervise the model's training in combination with a small percentage of

labeled data. Specifically, pseudo-labels are constructed by extracting one-hot labels from highly confident predictions on unlabeled data. These are then used as training targets in a supervised learning regime. This simplistic approach of utilizing unlabeled data has been combined with various methods to further improve the training efficiency. In Reference [1], authors demonstrated that setting a minimum number of labeled samples per training batch can be effective to reduce over-fitting due to noise accumulation on generated predictions. In addition, the use of a scalar temperature for scaling softmax output achieves a softer probability distribution over classes for the predictions and urges models to generate the correct pseudo-labels without suffering from over-confidence [11]. This temperature scaling approach can be highly beneficial in modern deep neural networks architectures, which have shown to suffer from over-confident predictions [11]. Supplementary, authors in [2] proposed MixMatch, which sharpens the prediction's distribution to further improve the generated pseudo-labels predictions. The sharpening process is performed by averaging the predictions' distribution of augmentation versions of the same unlabeled sample. Apart from self-training, alternative SSL approaches introduce a loss term, which is computed on unlabeled data, to encourage the model to generalize better to unseen data. Based on the objective of the loss term, we can classify these approaches in two categories: consistency regularization techniques—which are based on the principle that a classifier should produce the same class distribution for an unlabeled sample even after augmentation [31, 38]; and entropy minimization techniques—which aim to motivate the model to produce low-entropy (high-confident) predictions for all unlabeled data [8]. For a concise review and realistic evaluation of various deep learning based semi-supervised techniques, we refer interested readers to Reference [32].

A recent study [16] has questioned the soundness of the assumption that devices have well-annotated labels in a federated setting. Existing **semi-supervised federated learning (SSFL)** approaches, such as FedMatch [15] and FedSemi [26], have only recently started to be examined under the vision domain to exploit unlabeled data. FedMatch decomposes the parameters learned from labeled and unlabeled on-device data and uses an inter-client consistency loss to enforce consistency between the pseudo-labeling predictions made across multiple devices. In Reference [26], FedSemi adapts a mean teacher approach to harvest the unlabeled data and proposes an adaptive layer selection to reduce the communication cost during the training process. Apart from these methods, many studies consider different data distribution schemes, including sharing an unlabeled dataset across devices [14]. Last, it is important to note that recent works employ SSFL to address problems in healthcare domain, namely, electronic health records [13] and for problems like human activity recognition [39]. Nevertheless, none of the discussed approaches focuses on learning models for audio recognition tasks by utilizing devices' unlabeled audio samples.

3 BACKGROUND

In this section, we provide a brief overview of semi-supervised and federated learning paradigms, as they act as fundamental building blocks of our federated self-training approach for utilizing large-scale on-device unlabeled audio data in a federated setting.

3.1 Semi-supervised Learning

Given enough computational power and supervised data, deep neural networks have proven to achieve human-level performance on a wide variety of problems [21]. However, the curation of large-scale datasets is very costly and time-consuming, as it either requires crowd-sourcing or domain expertise, such as in the case of medical imaging. Likewise, for several practical problems, it is simply not possible to create a large enough labeled dataset (e.g., due to privacy issues) to learn a model of reasonable accuracy. In such cases, SSL algorithms offer a compelling alternative

to fully supervised methods for jointly learning from the fraction of labeled and a large number of unlabeled instances.

Specifically, SSL aims to solve the problem of learning with partially labeled data where the ratio of unlabeled training examples is usually much larger than that of the labeled ones. Formally, let $\mathcal{D}_L = \{(x_{l_i}, y_i)\}_{i=1}^{N_l}$ represent a set of labeled data, where N_l is the number of labeled data, x_{l_i} is an input instance, $y_i \in \{1, \dots, C\}$ is the corresponding label, and C is the number of label categories for the C -way multi-class classification problem. Besides, we have a set of unlabeled samples denoted as $\mathcal{D}_U = \{x_{u_i}\}_{i=1}^{N_u}$, where N_u is the number of unlabeled data. Let $p_\theta(y | x)$ be a neural network that is parameterized by weights θ that predicts softmax outputs $\hat{y}$ for a given input x . In the setting of semi-supervised learning, where in general $N_l \ll N_u$, we need to simultaneously minimize losses on both labeled and unlabeled data to learn the model's parameters θ . Specifically, our objective is to minimize the following loss function:

$$\mathcal{L}_\theta = \mathcal{L}_{s_\theta}(\mathcal{D}_L) + \mathcal{L}_{u_\theta}(\mathcal{D}_U), \quad (1)$$

where $\mathcal{L}_{s_\theta}(\mathcal{D}_L)$ and $\mathcal{L}_{u_\theta}(\mathcal{D}_U)$ are the loss terms from supervised and unsupervised learning, respectively.

The teacher-student self-training framework is a popular scheme to simultaneously learn from both labeled and unlabeled data. In this approach, we first use the available labeled data to train a good teacher model, which is then utilized to label any available unlabeled data. Consequently, both labeled and unlabeled data are used to jointly train a student model. In this way, the model assumes a dual role as a teacher and a student. In particular, as a student, it learns from the available data, while as a teacher, it generates targets to help the learning process of student. Since the model itself generates targets, they may very well be incorrect, thus, the learning experience of the student model depends solely on the ability of teacher model to generate high-quality targets [43].

3.2 Federated Learning

FL is a novel collaborative learning paradigm that aims to learn a single, global model from data stored on remote clients with no need to share their data with a central server. In particular, with the data residing on clients' devices, a subset of clients is selected to perform a number of local SGD steps on their data in parallel in each communication round. Upon completion, clients exchange their models' weights updates with the server, aiming to learn a unified global model by aggregating these updates. Formally, the goal of FL is typically to minimize the following objective function:

$$\min_{\theta} \mathcal{L}_\theta = \sum_{k=1}^K \gamma_k \mathcal{L}_k(\theta), \quad (2)$$

where $\mathcal{L}_k$ is the minimization function of the k th client and γ_k corresponds to the relative impact of the k th client to the construction of the global model. For the FedAvg algorithm, parameter γ_k is equal to the ratio of client's local data N_k over all training samples ($\gamma_k = \frac{N_k}{N}$).

Specifically, let $\mathcal{D} = \{(x_{l_i}, y_i)\}_{i=1}^N$ be a dataset of N labeled examples, similarly to the previously discussed dataset $\mathcal{D}_L$ in Section 3.1. Given K clients, $\mathcal{D}$ is decomposed into K sub-datasets $\mathcal{D}^k = \{(x_{l_i}, y_i)\}_{i=1}^{N_k}$ corresponding to each clients' privately held data. For an initial global model G , the r th communication round starts with server randomly selecting a portion q ($0 < q \leq K$) of clients to participate in the current training round. Afterwards, each client's local model receives the global parameters θ_r^G and performs supervised learning on their local dataset $\mathcal{D}^k$ to minimize $\mathcal{L}_k(\theta_r^k)$. Subsequently, G aggregates over locally updated parameters by performing $\theta_{r+1}^G \leftarrow \sum_{i=1}^q \frac{N_i}{N} \theta_r^i$. The presented circular training process, comprising model weights' exchanges between server and clients, repeats until θ^G converges after R rounds.

4 METHODOLOGY

In this section, we present our federated self-training learning approach, namely, FedSTAR, for audio recognition tasks. First, we provide a formal overview of the problem, which FedSTAR aims to solve. Next, we discuss the proposed self-training technique (i.e., pseudo-labeling with dynamic prediction confidence thresholding) in detail, followed by the presentation of our FedSTAR algorithm. Finally, we provide a thorough description of the self-supervised pre-training technique used to train a model as an initialization point for the FedSTAR approach.

4.1 Problem Formulation

We focus on the problem of *SSFL*, where labeled data are scarce across users' devices. At the same time, clients collectively hold a massive amount of unlabeled audio data. In addition, in a typical federated learning setting, the on-device data distribution depends on the profile of the users operating the devices. Thus, it is a common scenario for both labeled and unlabeled data to originate from the same data distribution. Based on the aforementioned assumption, with FedSTAR, we aim to utilize the available unlabeled data on clients and further improve the generalization of FL models, alleviating the need for clients to hold well-annotated data. In this way, we substantially decouple the amount of available labeled from the predictive power of acoustic models trained under federated settings.

Formally, under the setting of *SSFL*, each of the K clients holds a labeled set, $\mathcal{D}_L^k = \{(x_{l_i}, y_i)\}_{i=1}^{N_{l,k}}$ and an unlabeled set $\mathcal{D}_U^k = \{x_{u_i}\}_{i=1}^{N_{u,k}}$, where $N_k = N_{l,k} + N_{u,k}$ is the total number of data samples stored on the k th client and $N_{l,k} \ll N_{u,k}$. We desire to learn a global unified model G without clients sharing any of their local data, $\mathcal{D}_L^k$ and $\mathcal{D}_U^k$. To this end, our objective is to simultaneously minimize both supervised and unsupervised learning losses during each client's local training step on the r th round of the FL algorithm. Specifically, the minimization function, similar to the one presented in Equation (2), is:

$$\min_{\theta} \mathcal{L}_{\theta} = \sum_{k=1}^K \gamma_k \mathcal{L}_k(\theta) \text{ where } \mathcal{L}_k(\theta) = \mathcal{L}_{s_{\theta}}(\mathcal{D}_L^k) + \beta \mathcal{L}_{u_{\theta}}(\mathcal{D}_U^k). \quad (3)$$

Here, $\mathcal{L}_s(\mathcal{D}_L^k)$ is the loss terms from supervised learning on the labeled data held by the k th client, and $\mathcal{L}_u(\mathcal{D}_U^k)$ represents the loss term from unsupervised learning on the unlabeled data of the same client. We add the parameter β to control the effect of unlabeled data on the training procedure, while γ_k is the relative impact of the k th client on the construction of the global model G .

4.2 Self-training with Pseudo Labeling

Self-training via pseudo-labeling has been widely used in semi-supervised learning [40]. The objective of highly effective teacher-student self-training approaches is to train a teacher model, which supervises the learning process of a student model that learns from labeled and unlabeled data jointly. First, a teacher model is built with the available labeled data and afterwards this is exploited to make predictions for the unlabeled samples. Subsequently, the student model is trained on both labeled and predicted samples. We propose a self-training technique with a dynamic prediction confidence threshold to learn from the unlabeled audio data residing on the client's device, thus boosting the performance of models trained in federated settings with varying percentages of labeled examples. For audio classification tasks, to learn from the labeled datasets $\mathcal{D}_L^k$ across all participating clients, we apply cross-entropy loss as follows:

$$\mathcal{L}_s(\mathcal{D}_L^k) = -\frac{1}{N_{l,k}} \sum_{i=1}^{N_{l,k}} \sum_{j=1}^C y_i^j \log(f_i^{\theta^k}(x_{l_j})) = \mathcal{L}_{CE}(y, p_{\theta^k}(y | x_l)). \quad (4)$$

Next, to learn from unlabeled data, we generate pseudo-labels $\hat{y}$ for all available unlabeled data x_u on client k by performing:

$$\hat{y} = \Phi(z, T) = \arg \max_{i \in \{1, \dots, C\}} \left(\frac{e^{z_i/T}}{\sum_{j=1}^C e^{z_j/T}} \right), \quad (5)$$

where z_i are the logits produced for the input sample x_{u_i} by the k th client model p_{θ^k} before the softmax layer. In essence, Φ produces categorical labels for the given “soften” softmax values, in which temperature scaling is applied with a constant scalar temperature T . As the maximum of the softmax function remains unaltered, the predicted pseudo-label $\hat{y}$ is identical as if the original prediction (without scaling) for an unlabeled sample x_u was used; however, the prediction confidence is weakened. A dynamic threshold τ of confidence is proposed following a cosine schedule to discard low-confidence predictions when generating pseudo-labels. For the obtained pseudo-labels, we then perform standard cross-entropy minimization while using $\hat{y}$ as targets in the following manner:

$$\mathcal{L}_u(\mathcal{D}_U^k) = -\frac{1}{N_{u,k}} \sum_{i=1}^{N_{u,k}} \sum_{j=1}^C \hat{y}_i^j \log(f_i^{\theta^k}(x_{u_j})) = \mathcal{L}_{CE}(\hat{y}, p_{\theta^k}(x_u)). \quad (6)$$

Revising the initial minimization goal of FedSTAR expressed in Equation (3), we can now represent local models’ loss function on the r th round of the FL algorithm for the k th client as:

$$\mathcal{L}_k(\theta^k) = \mathcal{L}_{CE}(y, p_{\theta^k}(y | x_l)) + \beta \mathcal{L}_{CE}(\hat{y}, p_{\theta^k}(x_u)). \quad (7)$$

4.3 Federated Self-training

The objective of federated self-training is to create a teacher model on each client to exploit labeled data resident on clients’ devices, which will be used to predict labels for the unlabeled instances available in the device. As both labeled and unlabeled on-device samples originate from the same data distribution, a student model can be constructed on each client device by collectively training on labeled and pseudo-labeled data, the weights of whom will be returned to the server for aggregation. Under federated settings, however, a more complicated analysis is required, as clients’ local labeled data can be limited and can have a highly skewed distribution. In such settings, teacher models may produce inaccurate pseudo-label predictions, and student classifiers potentially amplify the mistakes further during training through using faulty pseudo-labels. To ensure the proper construction of pseudo-labels and guarantee that the student model will learn properly from unlabeled data, the confidence of the predictions is taken into consideration when generating pseudo-labels to discard any low-confidence predictions.

Concisely, in the proposed FedSTAR algorithm, the clients’ local update step is altered to learn from unlabeled datasets $\mathcal{D}_U^k$. As can be seen in Figure 2, a representative round r of FedSTAR starts with the distribution of global models’ weights θ^G to a randomly selected subset of q clients. On each client, equally sized batches l and u from the labeled and unlabeled sets are created, respectively. The model’s weights update is performed, as in Equation (7). First, the classical supervised categorical cross-entropy loss is minimized for batch l , as in Equation (4), to construct a teacher model, and afterwards, with this model pseudo-labels are produced using $\Phi(\cdot)$. With the creation of pseudo-labels $\hat{y}$, the unlabeled batch u is then treated as a labeled batch $u' = \{(u, \hat{y})\}$, in which the client’s model is further trained with standard cross-entropy minimization. It is important to note that we simultaneously optimize the cross-entropy loss on both l and u subsets by computing both losses before performing backpropagation to update the local models’ parameters. Last, the locally updated weights from all participating clients in the r th round are sent back to the server, where the global model weights are calculated as a weighted average over all the local weights updates.

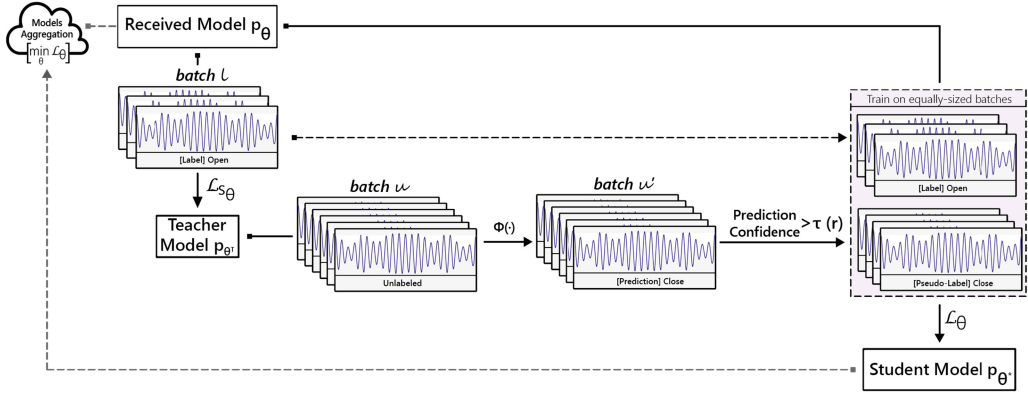


Fig. 2. On-device self-training based on pseudo-labeling in a federated setting for an audio recognition task shown for illustration purposes.

Since $N_l \ll N_u$ holds for all clients, given a sufficient number of participating rounds, unlabeled instances will be exposed to all the available labeled data. Additionally, we propose an *adaptive confidence thresholding* method to diminish unsatisfactory performance due to training on faulty pseudo-labels. In particular, in addition to using temperature scaling T to “soften” softmax output and generated confident predictions, we employ an increasing confidence threshold τ to discard low-confidence pseudo-labels during training following a cosine schedule. Cosine learning rate schedulers rely on the observation that we might not want to decrease the learning rate too drastically in the beginning, while we might want to “refine” our solution in the end using a very small learning rate. Along the same lines, with our cosine confidence thresholding, we allow clients to explore the locally stored unlabeled data, D_u^k , in the first few federated rounds, while considering only highly confident predictions in a later stage of the training procedure. While other methods could be explored for this purpose, such a study is outside the scope of the current work and we mainly focus on cosine scheduler, which has proven to work well empirically across a variety of tasks [27]. Further details and an overview of our approach for the semi-supervised training procedure can be found in Algorithm 1.

4.4 Self-supervised Pretraining Strategy

Self-supervised learning aims to learn useful representations from unlabeled data by tasking a model to solve an auxiliary task for which supervision can be acquired from the input itself. Given an unlabeled data $D = \{x\}_{m=1}^M$ and deep neural network $f_\theta(\cdot)$, the aim is to pre-train a model through solving a surrogate task, where labels y for the standard objective function (e.g., cross-entropy) are extracted automatically from x . The learned model is then utilized as a fixed feature extractor or as initialization for rapidly learning downstream tasks of interest. The fields of computer vision and natural language processing have seen tremendous progress in representation learning with deep networks in a self-learning manner, with no human intervention in the labeling process. Here, the prominent techniques for audio representation learning from unlabeled data include and audio-visual synchronization [20], contrastive learning [35], and other auxiliary tasks [37].

In our work, we propose to leverage self-supervised pre-training on the server side to improve training convergence of FedSTAR on client devices. Motivated by the fact that the server can often hold a large amount of unlabeled data and has enormous computational resources available, we

ALGORITHM 1: FedSTAR: Federated Self-training for Audio Recognition. In the algorithm, l and u are equally sized batches from on-device labeled and unlabeled samples, respectively. Scalar β controls the affect of unlabeled data in the training process, and η is the learning rate.

```

1: Server initialization of model  $G$  with model weights  $\theta_0^G$ 
2: for  $i = 1, \dots, R$  do
3:   Randomly select  $K$  clients to participate in round  $i$ 
4:   for each client  $k \in K$  in parallel do
5:      $\theta_i^k \leftarrow \theta_i^G$ 
6:      $\theta_{i+1}^k \leftarrow \text{ClientUpdate}(\theta_i^k)$ 
7:   end for
8:    $\theta_{i+1}^G \leftarrow \sum_{k=1}^K \frac{N_k}{N} \theta_{i+1}^k$ 
9: end for
10: procedure ClientUpdate( $\theta$ )
11:   for epoch  $e = 1, 2, \dots, E$  do
12:     for batch  $l \in \mathcal{D}_L$  and  $u \in \mathcal{D}_U$  do
13:        $\hat{y} \leftarrow \Phi(p_\theta(x_u), T)$ 
14:        $\theta \leftarrow \theta - \eta \nabla_\theta (\mathcal{L}_{CE}(y, p_\theta(y | x_l)) + \beta \cdot \mathcal{L}_{CE}(\hat{y}, p_\theta(x_u)))$ 
15:     end for
16:   end for
17: end procedure
    
```

employ contrastive learning for audio to develop a model that can be used as an effective starting point for federated learning instead of using randomly initialized weights. Specifically, pre-training is performed in a centralized setting with a separate publicly available dataset on the server side; thus, it can be done once and be used repeatedly for different downstream tasks. To the best of our knowledge, this is the first time self-supervised learning has been used to address the convergence of federated models with fewer training rounds efficiently.

Formally, we pre-train our model with contrastive learning [35] using FSD-50K [6] dataset. On a high level, the objective is to train a model to maximize the similarity between related audio segments while minimizing it for the rest. Similar samples are generated through stochastic sampling from the same audio clip, while other segments in a batch are treated as negatives. In particular, we use bilinear similarity formulation and pre-train our model with a batch size of 1,024 for 500 epochs. Moreover, we utilize a network architecture, as described in Section 5.2 as an encoder with the addition of a dense layer containing 256 hidden units on top, which is discarded after the pre-training stage. In this way by using a same architecture, we are able to draw proper conclusions for the effects of utilizing a pre-trained model as an initial global model and directly compare with the randomly initialized FedSTAR approach.

5 EXPERIMENTS

In this section, we conduct an extensive evaluation of our approach on publicly available datasets for various audio recognition tasks to determine the efficacy of FedSTAR in learning generalizable models under a variety of federated settings and label availability. First, the federated learning framework and datasets utilized for validation are presented, followed by a detailed description of the network architecture. Next, we introduce our experiments in centralized and fully supervised federated settings, which serve as a baseline for evaluating our approach. Finally, we provide a thorough evaluation of FedSTAR, which is structured in the form of several research questions.

Table 1. Key Details of the Datasets Used in Evaluation

Dataset	Task	Classes
Ambient Context [33]	Event classification	10
Speech Commands [41]	Keyword spotting	12
VoxForge [29]	Language identification	6

5.1 Datasets and Audio Pre-processing

We use publicly available datasets to evaluate our models on a range of audio recognition tasks. For all datasets, we use the suggested train/test split for comparability purposes. For ambient sound classification, we use the Ambient Acoustic Contexts dataset [33], in which sounds from 10 distinct events are present. For the keyword spotting task, we use the second version of the Speech Commands dataset [41], where the objective is to detect when a particular keyword is spoken out of a set of 12 target classes. Likewise, we use VoxForge [29] for the task of spoken language classification, which contains audio recordings in six languages—English, Spanish, French, German, Russian, and Italian. It is one of the largest available datasets for language identification problems; it is valuable for benchmarking the performance of the supervised FL model. We resampled the Ambient Acoustic Contexts samples from 48 kHz to 16 kHz to utilize the same sampling frequency across all our datasets samples. In Table 1, we present a description of each dataset.

5.2 Model Architecture and Optimization

The network architecture of our global model is inspired by Reference [37] with a key distinction that instead of batch normalization, we utilize group normalization [42] after each convolutional layer and employ a spatial dropout layer. We use log-Mel spectrograms as the model’s input, which we compute by applying a short-time Fourier transform on the one-second audio segment with a window size of 25 *ms* and a hop size equal to 10 *ms* to extract 64 Mel-spaced frequency bins for each window. To make an accurate prediction on an audio clip, we average over the predictions of non-overlapping segments of an entire audio clip. Our convolutional neural network architecture consists of four blocks. In each block, we perform two separate convolutions, one on the temporal and another on the frequency dimension, outputs of which we concatenate afterward to perform a joint 1×1 convolution. Using this scheme, the model can capture fine-grained features from each dimension and discover high-level features from their shared output. Furthermore, we apply L2 regularization with a rate of 0.0001 in each convolution layer and group normalization [42] after each layer. Between blocks, we utilize max-pooling to reduce the time-frequency dimensions by a factor of two and use a spatial dropout rate of 0.1 to avoid over-fitting. We apply ReLU as a non-linear activation function and use Adam optimizer with the default learning rate of 0.001 to minimize categorical cross-entropy.

To simulate a federated environment, we use the Flower framework [3] and utilize FedAvg [30] as an optimization algorithm to construct the global model from clients’ local updates. Additionally, a number of parameters were selected to control the federated settings of our self-training strategy fully. Those parameters are: (1) N - number of clients, (2) R - number of rounds, (3) q - clients’ participation percentage in each round, (4) E - number of local train steps per round, (5) σ - data distribution variance across clients, (6) L - dataset’s percentage to be used as labeled samples, (7) U - dataset’s percentage to be used as unlabeled samples (excluding $L\%$ of the data used as labeled), (8) β - influence of unlabeled data over training process, (9) T - temperature scaling parameter, and (10) τ - predictions confidence threshold. We employ uniform random sampling for the clients’ selection strategy, as other approaches for adequate clients election are outside the current work scope. Last, across all FedSTAR experiments, we fixed the temperature scaling

Table 2. Primary Experiment Parameters

Parameter Name	Variable	Range
Number of Clients	N	5–30
Number of Federated Rounds	R	1–100
Number of Local Train Steps	E	1–4
Clients' Participation Percentage	q	20% to 80%
Data Distribution Variance across Clients	σ	0% to 50%
Dataset's Labeled Percentage	L	3% to 100%
Dataset's Utilized Unlabeled Percentage	U	20% to 100%
Unlabeled Data Influence on Train Step	β	50%
Temperature Scaling	T	4
Confidence Threshold Percentage	τ	50% to 90%

parameter $T = 4$, while we set the confidence threshold τ to initialize from 0.5 and gradually increase to a maximum of 0.9 during training, following a cosine schedule. A description of the parameters used is presented in Table 2.

5.3 Baselines and Evaluation Strategy

In fully supervised federated experiments where the complete dataset is available, the labeled instances are randomly distributed across the available clients. Likewise, in experiments where the creation of a labeled subset from the original dataset is required ($L < 100\%$), we keep the dataset's initial class distribution ratio to avoid tempering with dataset characteristics. Afterward, the labeled subset is again randomly distributed across the available clients. With the σ parameter set to 25% and a random partitioning of labeled samples among clients, the labeled data distribution resembles a non-i.i.d. one. In contrast, an increase of available clients results in a highly skewed distribution. It is worth mentioning that even if the meaning of non-i.i.d. is generally straightforward, data can be non-i.i.d. in many ways. In our work, the term non-i.i.d. data distribution describes a distribution with both a label distribution skew and a quantity skew (data samples imbalance across clients). This type of data distribution is common across clients' data in federated settings. Each client frequently corresponds to a particular user (affecting the label distribution), and the application usage across clients can differ substantially (affecting the label distribution). For a concise taxonomy of non-i.i.d. data regimes, we refer our readers to Reference [17]. Additionally, in FedSTAR, the unlabeled subset consists of the dataset's remaining samples after extracting the provided labels. In such experiments, both the labeled and unlabeled subsets are dispensed at random over the available clients. Furthermore, for an accurate comparison between our experiments, we manage any randomness during the data partitioning and training procedures by passing a seed alongside the parameters presented in Table 2. In this way, we control the amount of data and the data instances that reside on each simulated client. Last, for a more rigorous evaluation, we perform three distinct trials (or runs, i.e., training a model from scratch) in each setting, and the average accuracy over all three runs is reported across the results of Sections 5.3 and 5.4.

To evaluate the FedSTAR, we first need to construct a high-quality *supervised* baseline both in centralized and federated environments. Therefore, we perform preliminary experiments in both centralized as well as fully supervised federated settings. We conduct initial experiments on all datasets in centralized settings where the models are trained until convergence to obtain the resulting accuracy on a test set, which is presented in the centralized row of Table 3. Following, we examine our model's performance in federated settings by adjusting the FL parameters to $R = 100$, $q = 80\%$, $E = 1$, and $\sigma = 25\%$. We vary the number of clients (N) while keeping the remaining federated parameters to the same as the earlier mentioned values, as N frequently fluctuates in

Table 3. Evaluation of Audio Recognition Models in Centralized and Fully supervised Federated Settings

Method		Speech Commands	Ambient Context	VoxForge
Centralized		96.54	73.03	79.60
Federated	$N = 5$	96.93	71.88	79.13
	$N = 10$	96.78	68.01	78.98
	$N = 15$	96.33	66.86	76.09
	$N = 30$	94.62	65.14	65.17

Average accuracy on test set over three distinct trials. Federated parameters are set to $q = 80\%$, $\sigma = 25\%$, $L = 100\%$, $E = 1$, $R = 100$.

real-life FL scenarios. Thus, we can explore how the federated model behaves as clients progressively increase and the available local data become yet more distributed, affecting the performance of FL [45]. The results for supervised FL are presented in the Federated row of Table 3. We note from results presented in Table 3 that the supervised federated models achieve comparable results in various cases to the models trained in a centralized setting across all three datasets. Moreover, the number of clients (N) has a clear effect on the model's performance. With an increase in N , we notice that the training process requires more training rounds (R) to converge as the quantity of local data for each client decreases. The obtained accuracy for a constant number of rounds deteriorates.

5.4 Results

5.4.1 Comparison against Fully Supervised Federated Approach under non-i.i.d. Settings. We first evaluate FedSTAR to determine the obtained improvements versus a fully supervised federated approach when a non-i.i.d. distribution is considered. This analysis helps in understanding *whether utilizing unlabeled instances that reside on clients' devices with FedSTAR can be beneficial for a model trained in federated settings and, if so, to which degree it improves the recognition rate*. To this end, we perform experiments on all three datasets for a diverse number of clients (N) where the percentage of available labeled instances is varied from 3% up to 50%. To clearly illustrate the performance gain of FedSTAR in comparison to the supervised FL regime, experiments with identically labeled subsets are conducted under fully supervised FL, where the unlabeled instances remained unexploited. Table 4 provides the accuracy scores on test sets averaged across three independent runs for the considered datasets to be robust against differences in weight initialization and optimization. For ease of comparison, we add the results column on fully supervised FL using entire labeled dataset ($L = 100\%$) in Table 4, as discussed earlier in Section 5.3.

In Table 4, we observe that FedSTAR can utilize unlabeled audio data to improve the model's performance across all datasets significantly. Consequently, we can conclude that FedSTAR can be applied in a federated environment with scarce labeled audio instances to boost the performance by learning from unlabeled data, independent of the audio recognition task. In particular, comparing the two rows for $L=3\%$, we note an increase of 13.28% in accuracy on average when using FedSTAR across the considered tasks.

While varying L , we note that the percentage gap between FedSTAR and the supervised federated counterparts shrinks as more labeled data are available across devices. In addition, with only 5% of labels available, we note that FedSTAR model's accuracy is within a reasonable range from the ones trained under fully supervised federated settings, where the complete dataset is available ($L = 100\%$). These two observations suggest that FedSTAR can be especially useful under extreme label scarcity scenarios, where a highly accurate model can be obtained through the exploitation of unlabeled data. Alternatively, FedSTAR could also be used in cases where sufficient labels are provided ($L = 50\%$) to slightly improve the resulting models' performance. An exception to the

Table 4. Performance Evaluation of FedSTAR

Dataset	Clients	Supervised (Federated)					FedSTAR			
		$L = 3\%$	$L = 5\%$	$L = 20\%$	$L = 50\%$	$L = 100\%$	$L = 3\%$	$L = 5\%$	$L = 20\%$	$L = 50\%$
Ambient Context	5	46.34	47.89	61.40	65.85	71.88	48.68	54.95	64.37	67.04
Speech Commands		81.12	87.97	92.35	94.66	96.93	87.41	90.01	94.17	94.85
VoxForge		54.55	56.41	61.65	70.37	79.13	63.92	67.80	69.09	67.08
Ambient Context	10	35.29	41.31	51.71	62.69	68.01	48.87	52.37	62.94	64.42
Speech Commands		67.75	83.80	92.12	94.02	96.78	86.82	90.33	94.09	94.18
VoxForge		56.14	54.73	60.48	62.41	78.98	59.87	64.35	69.38	63.27
Ambient Context	15	33.03	42.75	53.37	59.97	66.86	49.54	54.71	63.46	62.41
Speech Commands		62.98	72.84	92.14	93.14	96.33	86.82	89.33	93.16	93.39
VoxForge		54.26	54.37	57.11	60.29	76.09	55.82	57.96	67.66	61.66
Ambient Context	30	32.31	40.17	47.05	55.85	65.14	40.84	46.58	60.21	56.19
Speech Commands		33.78	44.21	84.94	92.21	94.62	83.88	88.19	92.92	92.62
VoxForge		50.32	54.33	55.19	57.56	65.17	54.81	56.18	63.83	56.66

Average accuracy over three distinct trials on test set. Detailed results are given in Table 9 of the Appendix. Federated parameters are set to $q = 80\%$, $\sigma = 25\%$, $\beta = 0.5$, $E = 1$, $R = 100$.

forementioned behavior is the case of FedSTAR experiments on VoxForge with $L = 50\%$, where we can see that the accuracy obtained after $R = 100$ rounds is inferior to the one achieved with identical federated settings and $L = 20\%$. This might be because the unlabeled subset was not yet exposed to all available labeled examples; thus, the model had not reached the learning plateau and may require more training rounds to converge. Despite this behavior, the performance of FedSTAR models is superior to the supervised federated models with the same amount of labeled data and reasonably close to supervised federated models trained on the entire labeled data.

While N increases and the labeled subset of each client shrinks (and, hence, we obtain an even higher non-i.i.d. distribution), we notice that the FedSTAR models' accuracy remains relatively unaffected, especially if we recollect that in FL experiments of Table 3 we noticed accuracy decays for a constant R as N rises. In particular, when $N = 30$ and the data distribution becomes highly skewed (or non-i.i.d.), we note a performance gap between FedSTAR and supervised FL that can reach up to 50% for the Speech Commands dataset. It means that the FedSTAR model can effectively utilize unlabeled audio data, even in highly distributed scenarios. This essentially means that the exploitation of large-scale unlabeled on-device instances can help create a more uniform distribution across devices and tackle the challenges introduced in federated settings from the "non-i.i.d.-ness" of data.

5.4.2 Effectiveness of FedSTAR across Diverse Federated Settings. In this subsection, we assess the efficacy of FedSTAR across a variety of federated settings. As presented in Section 5.4.1, the performance improvement by utilizing FedSTAR in comparison to the supervised FL scheme can vary across different federated settings. As federated settings' variability is a primary characteristic of a distributed environment, it is essential for our approach to be effective in distinct scenarios. To this end, we conduct further experiments on the Speech Command dataset, where the participation rate (q), the number of clients (N), the local train step (E), and the data distribution across clients (σ) are varied. We choose to investigate these four parameters, as they primarily vary in a real-life FL setting, and they can have a significant effect on model performance [17, 45].

Varying participation rate: With the device heterogeneity and computational resources significantly varying across devices in a federated environment, a participation rate of 100% is probably an unrealistic assumption for most pragmatic FL applications [17]. As clients' participation rate (q) can greatly influence the convergence rate of an FL model, we evaluate FedSTAR performance while varying the participation rate in each federated round. Therefore, this assessment helps us in understanding whether FedSTAR can retain the same level of effectiveness under low levels of clients participation. To this end, we conduct experiments with $N = 15$ for various clients' participation (q)

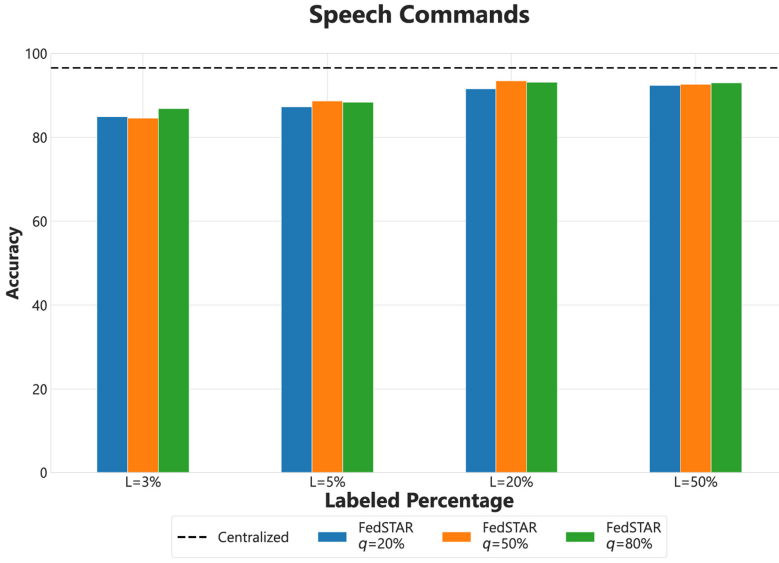


Fig. 3. Evaluation of FedSTAR performance under varying clients' participation rate. Federated parameters are set to $\sigma = 25\%$, $\beta = 0.5$, $R = 100$, $E = 1$, and $N = 15$. Average accuracy over three distinct trials is reported.

rates, starting from 20% up to 80%, under different percentages of labels availability on the Speech Commands dataset. Figure 3 provides obtained accuracy score on the test set; we observe that FedSTAR is able to effectively learn from the unlabeled instances residing on clients' devices under low levels of clients engagement, even if the available labeled samples are scarce. While there is a decrease in FedSTAR model's accuracy when the participation rate reduces, the reduction is no more than 2% for a given L . In particular, the reduction is eliminated when additional labeled instances are available.

Varying local train steps: Subsequently, we examine the effect of increasing the local train steps on the FedSTAR performance. As shown in Reference [25], a reduction in the communication costs can be achieved by increasing E at the expense of local models convergence, which can substantially affect the aggregation process. Thus, with this analysis, we aim to *understand whether FedSTAR models can retain their convergence rate when multiple local train steps are performed across clients' data to reduce the communication costs*. To this end, we perform experiments with various labeled percentages for 50 federated rounds ($R = 50$) and $N = 15$, while varying E from 1 to 4. From the results shown in Figure 4, we note that FedSTAR can effectively utilize the unlabeled instances, when the available labeled subset exceeds 3%, to avoid possible local models' divergence, resulting in a highly accurate aggregated (or global) model. However, for $L = 3\%$, we notice a declining trend as E increases, which could be originated from two reasons. Since the labeled data are scant for $L = 3\%$, the downwards trend on FedSTAR performance could be caused due to over-fitting, as the local models are extensively trained on a tiny labeled subset, when E increases. In addition, the absence of such a trend in higher label availability rates suggests that a sufficient amount of labeled data might be required for FedSTAR local models to converge. As FedSTAR uses unlabeled instances predictions to retrain the local models further, any faulty pseudo-labeled samples participating in the retraining step will increasingly intensify local models' divergence as E rises. Besides providing additional labels to avoid local models divergence, this behavior could be regulated by adjusting the confidence threshold of the predictions, τ , to a higher value so any initial faulty pseudo-labels would participate in the local SGD steps are discarded.

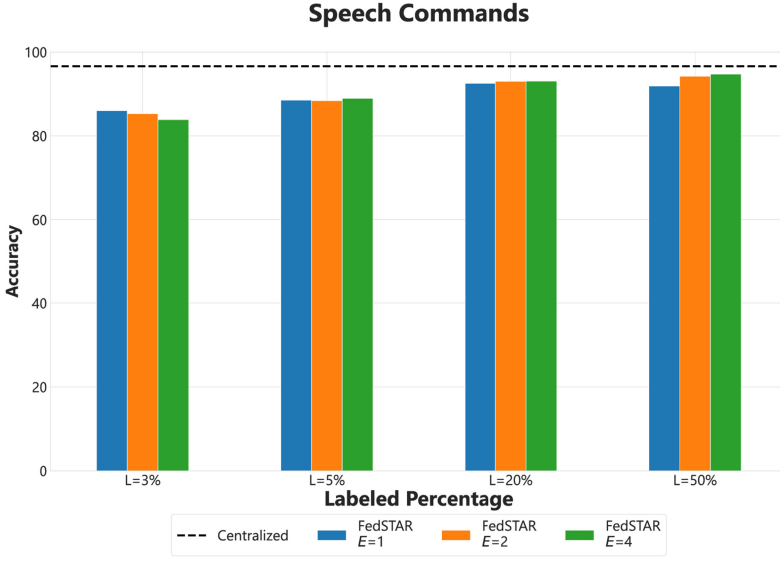


Fig. 4. Evaluation of FedSTAR performance against local train steps size. Federated parameters are set to $\sigma = 25\%$, $\beta = 0.5$, $R = 50$, $q = 80\%$, and $N = 15$. Average accuracy over three distinct trials is reported.

Varying number of clients: The number of clients is an important factor in the FL procedure, as it can have a significant impact on the data distribution, which has shown to affect the global model’s generalization [17, 45]. In particular, introducing additional clients to FL, the class distributions across clients can become highly skewed, as the data partitioning process is random. With this ablation study, we aim to answer *whether FedSTAR can retain the same level of effectiveness when the number of clients (and thus the non-i.i.d-ness of the class distribution) grows*. To this end, we present the performance of FedSTAR on Speech Commands dataset when we vary the number of clients from 5 up to 30, while setting $q = 80\%$, $\sigma = 25\%$, $R = 100\%$, and $E = 1$. The findings are shown in Figure 5, where we note that the number of clients has a relatively low impact on FedSTAR ability to utilize the available unlabeled audio data, as FedSTAR models’ performance follows a constant upward trend while we provide additional labels for any given N . In particular, comparing the results for $N = 5$ and $N = 30$, we observe that the FedSTAR models’ accuracy is notably close, especially for $L > 3\%$. This is in constant with fully supervised FL performance, as presented in Table 3, where models’ performance can drop more than 2% when varying N for the same dataset. Finally, it is important to note that FedSTAR model’s performance is close to the centralized baseline for both $L > 20$ and $L = 50$; thus, no noticeable improvement appears with the introduction of additional labels samples.

Varying class distribution across clients: Apart from the number of clients, the preferences of each client can substantially affect the nature of clients’ data distribution. For example, in a music tagging scenario, the type and quantity of data residing on a device are directly correlated to both user’s preference of a specific genre of music and the time users dedicated to the application. Such challenges introduce a highly non-i.i.d. data distribution, both in terms of labels distribution and quantity of data per client. Therefore, in this analysis, we aim to *understand the effect of highly non-i.i.d. distributions, both in terms of labels and data quantity distributions, on the effectiveness of FedSTAR to utilize on-device unlabeled data*. To this end, we execute experiments on the Speech Commands dataset with $N = 15$, $q = 80\%$, and $E = 1$ for $R = 100$, in which the partitioning of labeled data on clients followed a defined class availability distribution. We utilize a uniform distribution with a mean value of $\mu = 3$ and fluctuating variance σ_c from 0% to 50% as our class availability

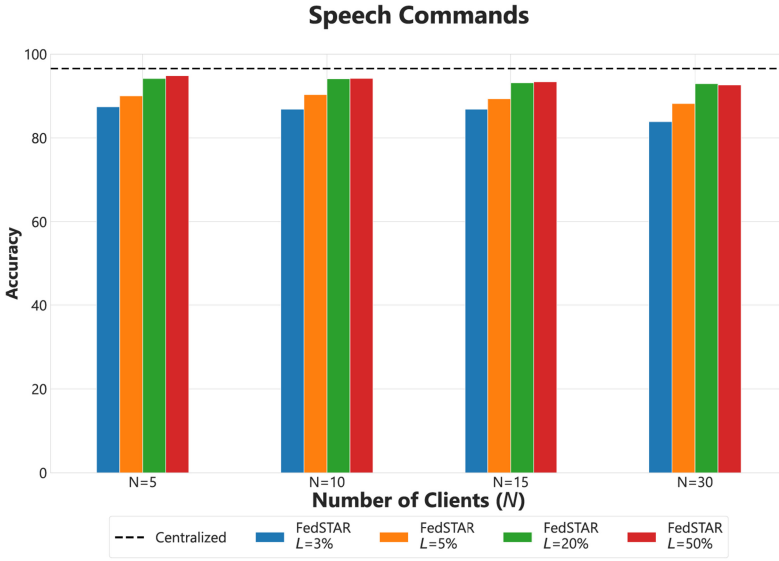


Fig. 5. Evaluation of FedSTAR performance under varying number of clients. Federated parameters are set to $\sigma = 25\%$, $\beta = 0.5$, $R = 100$, $q = 80\%$, and $E = 1$. Average accuracy over three distinct trials is reported.

distribution across clients. Since the total number of classes in the Speech Commands dataset is 12, we choose $\mu = 3$ for clients to access only a few labeled samples per class (on average, three classes). Thus, the on-device labeled data distribution resembles a realistic non-i.i.d. distribution. The client's preferences can affect both the type and the number of labeled samples described earlier for a music tagging application. It is important to note that the splitting of the unlabeled subset on clients followed a random distribution, with no assumption being made to distribute the label. Consequently, clients might have labeled samples from a specific subset of classes, yet unlabeled instances from all classes could be available. Such data distributions are frequent in pragmatic applications, where the domain knowledge is missing to perform the annotation process appropriately for all classes. For a rigorous evaluation, we perform identical experiments in terms of on-device labeled samples availability under fully supervised federated settings, where the unlabeled dataset remained unexploited.

From the results introduced in Table 5, we note that FedSTAR can effectively exploit the available on-device unlabeled instances to learn an accurate audio model under highly non-i.i.d. distributions. Comparing the FedSTAR performance with that of a fully supervised FL counterpart, we notice a substantial improvement in accuracy in most cases. In particular, for the case of $L \leq 3\%$, FedSTAR utilized on-device unlabeled examples to effectively train an audio model, whereas FL was unable to learn under such highly non-i.i.d. settings adequately. Additionally, we observe that the obtained accuracy gap across three distinct FedSTAR models (with σ_c of 0, 25, and 50 percent) for a given L is no larger than 4.8%. This behavior suggests that FedSTAR can maintain nearly the same level of effectiveness in exploiting on-device unlabeled data, irrespective of the skewness of data distribution on clients' end. Consequently, FedSTAR could be an effective solution to train an audio model under different federated settings, where the labeled data across clients experience a class distribution skewness and large-scale unlabeled audio samples from all classes are readily available on clients' devices.

Table 5. Performance Evaluation of Method against Variation of Class Availability across Clients

Class Distribution Characteristics		Supervised (Federated)				FedSTAR			
		$L = 3\%$	$L = 5\%$	$L = 20\%$	$L = 50\%$	$L = 3\%$	$L = 5\%$	$L = 20\%$	$L = 50\%$
$\mu = 3$	$\sigma_c = 0\%$	9.83	32.63	80.22	82.40	79.08	79.62	87.01	83.14
	$\sigma_c = 25\%$	10.54	23.97	75.41	83.61	79.05	84.15	86.52	85.05
	$\sigma_c = 50\%$	8.44	24.25	73.93	84.41	78.14	81.88	84.56	84.55

Class distribution has mean $\mu = 3$ and variance σ_c . Average accuracy over three distinct runs is reported on Speech Commands. Detailed results are given in Table 8 of the Appendix. Federated parameters are set to $\beta = 0.5$, $R = 100$, $N = 15$, $q = 80\%$, and $E = 1$.

5.4.3 Assessment of Utilizing Self-supervised Learning for Model Pre-training to Improve Training Convergence of FedSTAR. Our proposed self-training federated learning approach attains high performance on different audio recognition tasks by utilizing unlabeled data available on clients' end. However, in reality, a large volume of unlabeled instances from a different task or distribution might also be available on the centralized server. As servers often possess the computational power to effectively pre-train a model on a massive unlabeled dataset, a natural question arises, *whether leveraging self-supervised learning to pre-train a model as initialization for FedSTAR could improve the training convergence in federated settings with fewer rounds*. To this end, we perform experiments on all three datasets with $N = 15$ while using a model trained with a self-supervised pre-training strategy introduced in Section 4.4. We compare the obtained accuracy after 10 rounds of training ($R = 10$) when utilizing a self-supervised pre-trained model as an initial starting point for FedSTAR in contrast with a randomly initialized FedSTAR model trained for the same number of federated rounds. For a more rigorous evaluation, we vary labels availability from $L=3\%$ up to $L=100\%$ across all our datasets. The findings are presented in Figures 6(a), (c), and (e), where the average accuracy over three distinct trials is reported. Furthermore, the average train loss for the case of $L = 50\%$ in the first 10 federated rounds ($R = 10$) is also reported in Figures 6(b), (d), and (f). We choose to report the average train loss for the case of $L = 50\%$, since we previously observed from Table 4 that FedSTAR models might require additional rounds to utilize unlabeled data effectively. Thus, we can demonstrate that utilizing a pre-trained model as initialization for FedSTAR can significantly boost training convergence.

From Figures 6(a), (c), and (e), we note that the utilization of a pre-trained model leads to higher accuracy within 10 rounds in almost all cases, suggesting that it was able to perform finer pseudo-labels predictions and accelerate the model's convergence. In particular, for the Ambient Context dataset, where the amount of available labeled instances per client is tiny (approximately 13 labeled samples per client for $L = 5\%$ and $N = 15$), we observe a substantial difference between the pre-trained and randomly initialized FedSTAR approaches. This behavior suggests that in cases where the amount of labels is exceedingly sparse, utilizing self-supervised learning via model pre-training can significantly shorten the federated rounds needed for convergence. The beneficial role of the model pre-training on FedSTAR can be observed in the train loss gap between the pre-trained and the randomly initialized versions of FedSTAR after 10 rounds, presented in Figures 6(b), (d), and (f).

5.4.4 Effectiveness of FedSTAR under Varying Amount of Unlabeled Data. As of now, we have assumed that unlabeled data is largely available across clients. However, it is intriguing to investigate the scenario where both the amount of labeled and unlabeled data varies. In this way, we could simulate two pragmatic scenarios: First, an abundant volume of unlabeled instances generated by clients devices (e.g., numerous IoT devices constantly monitoring the surrounding environment); and second, relatively small amount of unlabeled audio samples available (e.g., medical audio examples, where both obtaining and labeling data is expensive). In addition, the

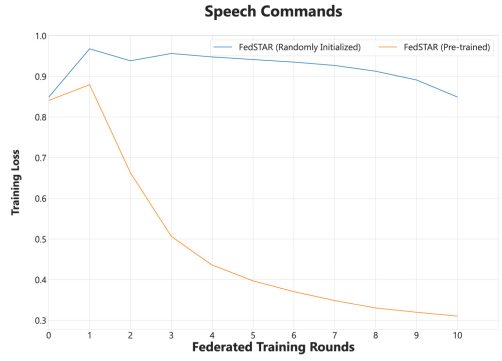
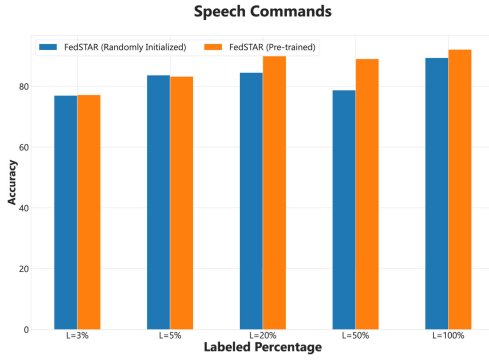
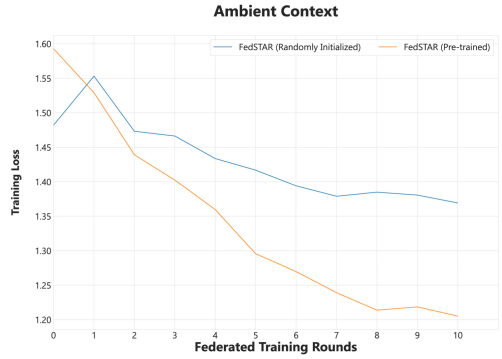
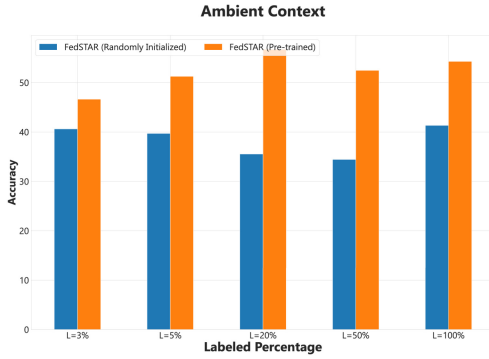
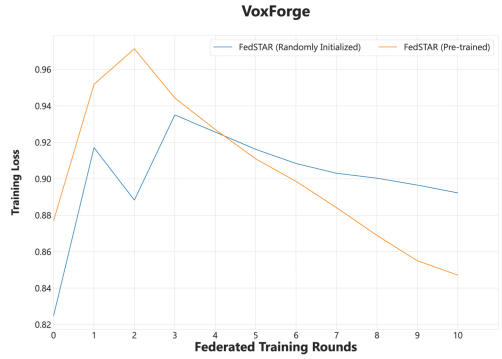
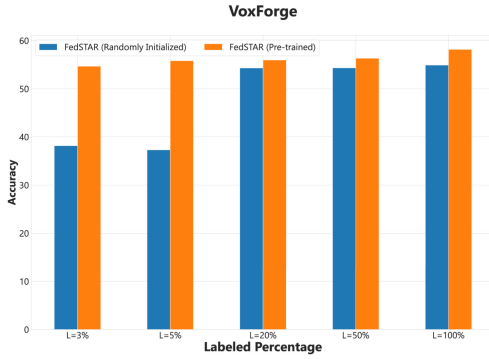
(a) Accuracy obtained after 10 rounds for $N=15$ on Speech Commands.(b) Train loss comparison for $N=15$ on Speech Commands.(c) Accuracy obtained after 10 rounds for $N=15$ on Ambient Context.(d) Train loss comparison for $N=15$ on Ambient Context.(e) Accuracy obtained after 10 rounds for $N=15$ on VoxForge.(f) Train loss comparison for $N=15$ on VoxForge.

Fig. 6. Self-supervised learning improves training convergence in federated setting. Federated parameters are set to $q = 80\%$, $\sigma = 25\%$, $\beta = 0.5$, $E = 1$. Average accuracy on testset over three distinct trials is reported.

restriction of available unlabeled on-device data could be originated from the often-limited storage capabilities of devices participating in the distributed machine learning paradigms. Thus, we aim to understand *the effect of unlabeled data availability on the FedSTAR efficiency to improve FL models' performance as well as the impact of utilizing pre-trained model with self-supervised.*

Table 6. Performance Evaluation of FedSTAR when Varying Both Labeled and Unlabeled Datasets

Labeled Percentage	FedSTAR (Randomly Initialized)				FedSTAR (SSL Pretrained)			
	$U = 20\%$	$U = 50\%$	$U = 80\%$	$U = 100\%$	$U = 20\%$	$U = 50\%$	$U = 80\%$	$U = 100\%$
$L = 3\%$	84.13	85.40	86.63	86.82	84.52	85.17	85.43	86.46
$L = 5\%$	87.47	88.52	88.90	89.33	88.07	88.28	87.73	88.98
$L = 20\%$	90.06	92.24	93.07	93.15	92.44	93.67	93.98	94.13
$L = 50\%$	87.76	92.26	94.18	93.38	90.70	93.83	94.76	95.54

Average accuracy over three distinct runs is reported on Speech Commands. Detailed results are given in Table 7 of the Appendix. Federated parameters are set to $q = 80\%$, $\sigma = 25\%$, $\beta = 0.5$, $R = 100$, $E = 1$, $N = 15$.

Consequently, we perform experiments on the Speech Commands dataset with $N = 15$, while varying the labeled subset from 3% up to 50% and the unlabeled dataset from 20% up to 100%. The obtained accuracy scores for both pre-trained and randomly initialized FedSTAR models are presented in Table 6.

As we see in Table 6, the availability of unlabeled data can affect the FedSTAR models performance. In particular, when the amount of both labeled and unlabeled instances is limited ($L \leq 5$ and $U \leq 50$), the obtained accuracy for both the randomly initialized and the pre-trained FedSTAR models is similar. However, as the amount of labeled data increases ($L > 5$), we notice a performance improvement of the pre-trained over the standard FedSTAR approach, in case of $L = 50$ and $U = 20$ results an accuracy difference of 3%. This behavior suggests that the utilization of on-device unlabeled data for the pre-trained FedSTAR is superior to that of the randomly initialized FedSTAR when sufficient labeled samples are provided. In addition, comparing results of both FedSTAR approaches in Table 6 with those for supervised federated with $N = 15$ in Table 4, we observe an accuracy improvement for both pre-trained and randomly initialized FedSTAR models over their supervised counterparts for $L < 20$. For higher labels availability ($L > 20$), the performance gap between the pre-trained FedSTAR models and the fully supervised federated alternatives is still prevalent, even for small volumes of unlabeled data ($U = 20$). From this, we can deduce that FedSTAR can effectively utilize unlabeled instances to improve the performance of audio models, even when the availability of on-device unlabeled samples is insufficient.

6 CONCLUSIONS AND FUTURE WORK

We study the pragmatic problem of semi-supervised federated learning for audio recognition tasks. In the distributed scenario, clients' well-annotated audio examples are deficient due to the prohibitive cost of annotation, users with little to no incentives to label their data, and notably for various important tasks, the domain knowledge that is missing to perform the annotation process appropriately. Conversely, large-scale unlabeled audio data are readily available on clients' devices. To address the lack of labeled data for learning on-device models, we present a novel self-training strategy based on pseudo-labeling to exploit on-device unlabeled audio data and boost the generalization of models trained in federated settings. Despite its simplicity, we demonstrate that our approach, FedSTAR, is highly feasible for semi-supervised learning on various audio recognition tasks within different federated settings and labels availability. We exhaustively evaluate FedSTAR on several publicly available datasets while comparing its performance with fully supervised federated and traditional centralized counterparts. The models' accuracy we achieve is consistently superior to fully supervised federated settings under the same labels availability. In many cases, FedSTAR results are comparable to fully supervised federated settings, where the complete dataset with labels was utilized. Furthermore, FedSTAR can retain the same level of effectiveness on utilizing unlabeled instances, irrespective of the amount of labels available on clients. In addition, FedSTAR can

significantly improve the model's performance in settings where on-device labeled samples from only a subset of classes are present, while the unlabeled instances contain examples from all classes. By utilizing on-device unlabeled samples from all classes, the data distribution across devices becomes more uniform; thus, the local models' learning objectives converge. This non-i.i.d data distribution setting is frequent in pragmatic scenarios, where the expertise is missing to annotate samples from all available classes, e.g., physiological signals in the medical domain. Finally, we demonstrate that self-supervised pre-trained models can significantly improve training convergence in federated settings with fewer rounds when used as model initialization for federated training instead of randomly initialized weights.

Despite the wide applicability, as FedSTAR is based on self-training, it is still relying on a few well-annotated samples across all devices to properly exploit any additional unlabeled data. Without such labeled samples, the utilization of unlabeled samples through FedSTAR might bring undesirable results. In reality, however, such a limitation can be lifted by requesting from users to annotate two to three samples, which are inexpensive to acquire. With the number of devices in an FL network usually ranging from hundreds to even thousands of devices, this process will provide a sufficient labeled subset. This can be used to train model in conjunction with the massively available unlabeled data using FedSTAR to acquire a highly accurate model. Furthermore, inherent noise, which originates from the audio signal, is an additional challenge that can limit the applicability of FedSTAR in real-life applications. Depending on the type and the amount of noise, this could affect the performance of FedSTAR, making the exploitation of unlabeled samples counter-productive. In such cases, there is a range of methods available that can be introduced as a preprocessing step in the learning procedure to mitigate or denoise the signal with minimal effort.

In this work, we provided a federated self-training scheme to learn audio recognition models through a few on-device labeled audio data. In the Internet of Things era, this approach could be employed in a variety of applications, such as home automation, autonomous driving, the health-care domain, and smart wearable technologies. In particular, we believe that federated self-training is of immense value for learning generalizable audio models in settings, where labeled data are challenging to acquire. However, unlabeled data are available in vast quantities. We hope that the presented perspective of federated self-training inspires the development of additional approaches, specifically those combining semi-supervised learning and federated learning in an asynchronous fashion. Likewise, combining federated self-training with appropriate client selection techniques is another crucial area of improvement that will further improve the performance of deep models in federated learning scenarios. Finally, evaluation in a real-world setting (i.e., federate learning involving real devices) is of major importance to further understand the aspects that require improvements concerning statistical and system heterogeneities, energy, and labeled data requirements in the federated learning setting.

APPENDIX

Table 7. Performance Evaluation of FedSTAR when Varying Both Labeled and Unlabeled Datasets

Labeled Percentage	FedSTAR (Randomly Initialized)				FedSTAR (Pre-Trained)			
	$U = 20\%$	$U = 50\%$	$U = 80\%$	$U = 100\%$	$U = 20\%$	$U = 50\%$	$U = 80\%$	$U = 100\%$
$L = 3\%$	84.13 ± 0.004	85.40 ± 0.008	86.63 ± 0.002	86.82 ± 0.020	84.52 ± 0.001	85.17 ± 0.001	85.43 ± 0.001	86.46 ± 0.006
$L = 5\%$	87.47 ± 0.001	88.52 ± 0.005	88.90 ± 0.001	89.33 ± 0.007	88.07 ± 0.004	88.28 ± 0.002	87.73 ± 0.001	89.98 ± 0.002
$L = 20\%$	90.06 ± 0.003	92.24 ± 0.012	93.07 ± 0.001	93.15 ± 0.011	92.44 ± 0.001	93.67 ± 0.005	93.98 ± 0.003	94.13 ± 0.001
$L = 50\%$	87.76 ± 0.003	92.26 ± 0.005	94.18 ± 0.001	93.38 ± 0.001	90.70 ± 0.001	93.83 ± 0.003	94.76 ± 0.001	95.54 ± 0.007

Average accuracy over 3 distinct runs is reported on Speech Commands, including variance across experiments.

Federated parameters are set to $q = 80\%$, $\sigma = 25\%$, $\beta = 0.5$, $R = 100$, $E = 1$, $N = 15$.

Table 8. Performance Evaluation of Method against Variation of Class Availability Across Clients

Class Distribution Characteristics		Supervised (Federated)				FedSTAR			
		$L = 3\%$	$L = 5\%$	$L = 20\%$	$L = 50\%$	$L = 3\%$	$L = 5\%$	$L = 20\%$	$L = 50\%$
$\mu = 3$	$\sigma_c = 0\%$	9.83 ± 0.017	32.63 ± 0.097	80.22 ± 0.056	82.40 ± 0.048	79.08 ± 0.026	79.62 ± 0.034	87.01 ± 0.028	83.14 ± 0.069
	$\sigma_c = 25\%$	10.54 ± 0.016	23.97 ± 0.139	75.41 ± 0.055	83.61 ± 0.046	79.05 ± 0.052	84.15 ± 0.013	86.52 ± 0.032	85.05 ± 0.051
	$\sigma_c = 50\%$	8.44 ± 0.001	24.25 ± 0.140	73.93 ± 0.044	84.41 ± 0.043	78.14 ± 0.021	81.88 ± 0.031	84.56 ± 0.041	84.55 ± 0.055

Class distribution has mean $\mu = 3$ and variance σ_c . Average accuracy over 3 distinct runs is reported on Speech Commands, including variance across experiments. Federated parameters are set to $\beta = 0.5$, $R = 100$, $N = 15$, $q = 80\%$ and $E = 1$.

Table 9. Performance Evaluation of FedSTAR

Dataset	Clients	Supervised (Federated)				FedSTAR			
		$L = 3\%$	$L = 5\%$	$L = 20\%$	$L = 50\%$	$L = 3\%$	$L = 5\%$	$L = 20\%$	$L = 50\%$
Ambient Context	5	46.34 ± 0.009	47.89 ± 0.056	61.40 ± 0.001	65.85 ± 0.021	48.68 ± 0.004	54.95 ± 0.026	64.37 ± 0.012	67.04 ± 0.010
Speech Commands		81.12 ± 0.037	87.97 ± 0.047	92.35 ± 0.030	94.66 ± 0.012	87.41 ± 0.007	90.01 ± 0.001	94.17 ± 0.003	94.85 ± 0.001
VoxForge		54.55 ± 0.009	56.41 ± 0.021	61.65 ± 0.005	70.37 ± 0.021	63.92 ± 0.016	67.80 ± 0.018	69.09 ± 0.013	67.08 ± 0.016
Ambient Context	10	35.29 ± 0.006	41.31 ± 0.012	51.71 ± 0.009	62.69 ± 0.018	48.87 ± 0.004	52.37 ± 0.018	62.94 ± 0.024	64.42 ± 0.006
Speech Commands		67.75 ± 0.001	83.80 ± 0.029	92.12 ± 0.087	94.02 ± 0.036	86.82 ± 0.006	90.33 ± 0.007	94.09 ± 0.002	94.18 ± 0.006
VoxForge		56.14 ± 0.020	54.73 ± 0.001	60.48 ± 0.033	62.41 ± 0.014	59.87 ± 0.024	64.35 ± 0.003	69.38 ± 0.016	63.27 ± 0.032
Ambient Context	15	33.03 ± 0.002	42.75 ± 0.007	53.37 ± 0.004	59.97 ± 0.004	49.54 ± 0.005	54.71 ± 0.022	63.46 ± 0.004	62.41 ± 0.006
Speech Commands		62.98 ± 0.003	72.84 ± 0.001	92.14 ± 0.003	93.14 ± 0.004	86.82 ± 0.006	89.33 ± 0.002	93.16 ± 0.001	93.39 ± 0.007
VoxForge		54.26 ± 0.002	54.37 ± 0.009	57.11 ± 0.031	60.29 ± 0.001	55.82 ± 0.011	57.96 ± 0.025	67.66 ± 0.004	61.66 ± 0.007
Ambient Context	30	32.31 ± 0.004	40.17 ± 0.001	47.05 ± 0.001	55.85 ± 0.002	40.84 ± 0.041	46.58 ± 0.013	60.21 ± 0.013	56.19 ± 0.009
Speech Commands		33.78 ± 0.012	44.21 ± 0.016	84.94 ± 0.012	92.21 ± 0.008	83.88 ± 0.001	88.19 ± 0.005	92.92 ± 0.005	92.62 ± 0.007
VoxForge		50.32 ± 0.009	54.33 ± 0.015	55.19 ± 0.011	57.56 ± 0.002	54.81 ± 0.001	56.18 ± 0.005	63.83 ± 0.009	56.66 ± 0.009

Federated parameters are set to $q = 80\%$, $\sigma = 25\%$, $\beta = 0.5$, $E = 1$, $R = 100$.

Average accuracy over 3 distinct trials on test set is reported, including variance across experiments.

REFERENCES

- [1] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks (IJCNN)*. 1–8. DOI : <https://doi.org/10.1109/IJCNN48605.2020.9207304>
- [2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. arXiv:1905.02249 [cs.LG].
- [3] Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchu Qiu, Titouan Parcollet, and Nicholas D. Lane. 2020. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390* (2020).
- [4] Justin Chan, Thomas Rea, Shyamnath Gollakota, and Jacob E. Sunshine. 2019. Contactless cardiac arrest detection using smart devices. *NPJ Dig. Med.* 2, 1 (2019), 1–8.
- [5] Pasquale Foggia, Nicolai Petkov, Alessia Sagge, Nicola Strisciuglio, and Mario Vento. 2016. Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE Trans. Intell. Transport. Syst.* 17, 1 (2016), 279–288. DOI : <https://doi.org/10.1109/TITS.2015.2470216>
- [6] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2020. FSD50k: An open dataset of human-labeled sound events. *arXiv preprint arXiv:2010.00475* (2020).
- [7] Yan Gao, Titouan Parcollet, Javier Fernandez-Marques, Pedro P. B. de Gusmao, Daniel J. Beutel, and Nicholas D. Lane. 2021. End-to-end Speech Recognition from Federated Acoustic Models. arXiv:2104.14297 [cs.SD].
- [8] Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. In *17th International Conference on Neural Information Processing Systems (NIPS'04)*. The MIT Press, Cambridge, MA, 529–536.
- [9] Andrew Hard, Kurt Partridge, Cameron Nguyen, Niranjana Subrahmanya, Aishanee Shah, Pai Zhu, Ignacio Lopez Moreno, and Rajiv Mathews. 2020. Training Keyword Spotting Models on Non-IID Data with Federated Learning. arXiv:2005.10406 [eess.AS].
- [10] Andrew Hard, Kanishka Rao, Rajiv Mathews, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. Federated learning for mobile keyboard prediction. *CoRR* abs/1811.03604 (2018).
- [11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [stat.ML].
- [12] Hossein Hosseini, Sungrack Yun, Hyunsin Park, Christos Louizos, Joseph Soriaga, and Max Welling. 2020. Federated Learning of User Authentication Models. arXiv:2007.04618 [cs.LG].
- [13] Li Huang and Dianbo Liu. 2019. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *CoRR* abs/1903.09296 (2019).
- [14] Sohei Itahara, Takayuki Nishio, Yusuke Koda, Masahiro Morikura, and Koji Yamamoto. 2021. Distillation-based Semi-supervised Federated Learning for Communication-efficient Collaborative Training with Non-IID Private Data. arXiv:2008.06180 [cs.DC].
- [15] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. 2020. Federated Semi-supervised Learning with Inter-client Consistency. arXiv:2006.12097 [cs.LG].
- [16] Yilun Jin, Xiguang Wei, Yang Liu, and Qiang Yang. 2020. Towards Utilizing Unlabeled Data in Federated Learning: A Survey and Prospective. arXiv:2002.11545 [cs.LG].
- [17] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badi Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021. Advances and Open Problems in Federated Learning. arXiv:1912.04977 [cs.LG].
- [18] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Noboru Harada, and Keisuke Imoto. 2019. ToyADMOS: A Dataset of Miniature-machine Operating Sounds for Anomalous Sound Detection. arXiv:1908.03299 [eess.AS].
- [19] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2017. Federated Learning: Strategies for Improving Communication Efficiency. arXiv:1610.05492 [cs.LG].
- [20] Bruno Korbar, Du Tran, and Lorenzo Torresani. 2018. Cooperative learning of audio and video models from self-supervised synchronization. In *32nd International Conference on Neural Information Processing Systems*. 7774–7785.
- [21] Brenden M. Lake, Tomer D. Ullman, and Joshua B. Tenenbaum and Samuel J. Gershman. 2016. Building machines that learn and think like people. *CoRR* abs/1604.00289 (2016).
- [22] Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, Vol. 3.

- [23] David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau. 2019. Federated Learning for Keyword Spotting. arXiv:[1810.05512](#) [eess.AS].
- [24] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2019. Federated learning: Challenges, methods, and future directions. *CoRR* abs/1908.07873 (2019).
- [25] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated Optimization in Heterogeneous Networks. arXiv:[1812.06127](#) [cs.LG].
- [26] Zewei Long, Liwei Che, Yaqing Wang, Muchao Ye, Junyu Luo, Jinze Wu, Houping Xiao, and Fenglong Ma. 2020. FedSemi: An Adaptive Federated Semi-supervised Learning Framework. arXiv:[2012.03292](#) [cs.LG].
- [27] Ilya Loshchilov and Frank Hutter. 2016. SGDR: Stochastic gradient descent with restarts. *CoRR* abs/1608.03983 (2016).
- [28] Oisín Mac Aodha, Rory Gibb, Kate E. Barlow, Ella Browning, Michael Firman, Robin Freeman, Briana Harder, Libby Kinsey, Gary R. Mead, Stuart E. Newson, Ivan Pandourski, Stuart Parsons, Jon Russ, Abigail Szodoray-Paradi, Farkas Szodoray-Paradi, Elena Tilova, Mark Girolami, Gabriel Brostow, and Kate E. Jones. 2018. Bat detective-deep learning tools for bat acoustic signal detection. *PLOS Computat. Biol.* 14, 3 (03 2018), 1–19. DOI : <https://doi.org/10.1371/journal.pcbi.1005995>
- [29] Ken MacLean. 2018. Voxforge. *Ken MacLean*. Retrieved from <http://www.voxforge.org/home>. [Acedido em 2012] (2018).
- [30] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-efficient Learning of Deep Networks from Decentralized Data. arXiv:[1602.05629](#) [cs.LG].
- [31] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-supervised Learning. arXiv:[1704.03976](#) [stat.ML].
- [32] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D. Cubuk, and Ian J. Goodfellow. 2019. Realistic Evaluation of Deep Semi-supervised Learning Algorithms. arXiv:[1804.09170](#) [cs.LG].
- [33] Chunjong Park, Chulhong Min, Sourav Bhattacharya, and Fahim Kawsar. 2020. Augmenting conversational agents with ambient acoustic contexts. In *22nd International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCT'20)*. Association for Computing Machinery, New York, NY. DOI : <https://doi.org/10.1145/3379503.3403535>
- [34] Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. 2019. Federated learning for emoji prediction in a mobile keyboard. *CoRR* abs/1906.04329 (2019).
- [35] Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3875–3879.
- [36] Dan Stowell, Yannis Stylianou, Mike Wood, Hanna Pamula, and Hervé Glotin. 2018. Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge. *CoRR* abs/1807.05812 (2018).
- [37] Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, and Dominik Roblek. 2019. Self-supervised audio representation learning for mobile devices. *arXiv preprint arXiv:1905.11796* (2019).
- [38] Antti Tarvainen and Harri Valpola. 2018. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv:[1703.01780](#) [cs.NE].
- [39] Bram van Berlo, Aaqib Saeed, and Tanir Ozcelebi. 2020. Towards federated unsupervised representation learning. In *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*. Association for Computing Machinery, 31–36. <https://doi.org/10.1145/3378679.3394530>
- [40] Jesper E. van Engelen and H. Hoos. 2019. A survey on semi-supervised learning. *Mach. Learn.* 109 (2019), 373–440.
- [41] Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *CoRR* abs/1804.03209 (2018).
- [42] Yuxin Wu and Kaiming He. 2018. Group Normalization. arXiv:[1803.08494](#) [cs.CV].
- [43] Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Self-training with noisy student improves imagenet classification. *CoRR* abs/1911.04252 (2019).
- [44] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. 2018. Applied Federated Learning: Improving Google Keyboard Query Suggestions. arXiv:[1812.02903](#) [cs.LG].
- [45] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. 2018. Federated Learning with Non-IID Data. arXiv:[1806.00582](#) [cs.LG].
- [46] Xiaojin Zhu and Andrew Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3, 1 (2009), 1–130. DOI : [10.2200/s00196ed1v01y200906aim006](https://doi.org/10.2200/s00196ed1v01y200906aim006)

Received 14 July 2021; revised 5 February 2022; accepted 19 February 2022