FEDERATED SELF-TRAINING FOR DATA-EFFICIENT AUDIO RECOGNITION

Vasileios Tsouvalas[1], *Aaqib Saeed*[2], *Tanir Ozcelebi*[1]

[1]Eindhoven University of Technology, Eindhoven, The Netherlands
[2]Philips Research, Eindhoven, The Netherlands

ABSTRACT

Federated learning is a distributed machine learning paradigm dealing with decentralized and personal datasets. Since data reside on devices like smartphones, 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. Existing federated learning approaches largely focus on supervised learning without harnessing the unlabeled data. Here, we study the problem of semi-supervised learning of audio models in conjunction with federated learning. We propose `FedSTAR` , a self-training approach to exploit large-scale on-device unlabeled data to improve the generalization of audio recognition models. We conduct experiments on diverse public audio classification datasets and investigate the performance of our models under varying percentages of labeled data and show that with as little as 3% labeled data, `FedSTAR` on average can improve the recognition rate by 13.28% compared to the fully-supervised federated model.

Index Terms— federated learning, semi-supervised learning, deep learning, audio classification, sound recognition

1. INTRODUCTION

Federated Learning (FL) has been attracting growing attention thanks to its unique characteristic of collaboratively training machine learning models without sharing local data and compromising users' privacy [1]. The most popular paradigm in FL is the Federated Averaging (FedAvg) algorithm [2], where minimal updates to the models are performed entirely on-device and communicated to the central server, which aggregates updates to produce a unified global model. This strategy has been applied on a wide range of tasks [3, 4, 5, 6] and recently performing federated training of acoustic models has attracted considerable attention [3, 4, 7, 8]. 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, is in most pragmatic cases highly unrealistic.

In reality, on-device data is largely unlabeled and constantly expanding in size. Due to the prohibitive cost of annotation, users have little to no incentives to label the data. Notably, for various important tasks, the domain knowledge is missing to perform the annotation process appropriately that leaves most of the data residing on devices to remain unlabeled. This is especially accurate when considering the utilization of audio data to perform various recognition tasks, which has several real-world applications [9, 3, 10]. However, in the majority of these applications, there is no straightforward manner for the annotation process [11]. This leads to a novel FL problem, namely *semi-supervised federated learning*, where users' devices collectively hold a massive amount of unlabeled audio samples and only a fraction of labeled audio examples.

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 [12, 13]. In self-training or pseudo-labeling, we use the prediction on unlabeled data to supervise the model's training in combination with a small percentage of labeled data [14]. This simplistic yet effective approach has been shown to achieve great results in prior work [15, 16]. Existing semi-supervised federated learning (SSFL) approaches have only recently started to be examined under the vision domain to exploit unlabeled data [17, 18]. FedMatch [17] uses an inter-client consistency loss to enforce consistency between the pseudo-labeling predictions made across multiple devices. In [18], FedSemi adapts a mean teacher approach to harvest the unlabeled data during the training process. Nevertheless, none of the discussed approaches focuses on learning models for audio recognition tasks by utilizing devices' unlabeled audio samples and introduce additional communication overhead to utilize the available on-device unlabeled data.

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. `FedSTAR` aim to improve the generalization of federated models on a wide range of audio recognition tasks under a pragmatic scenario, where scarcity of labels and the non-i.i.d. nature of data across clients poses a significant challenge for learning useful models. To the best of our knowledge `FedSTAR` is the first FL approach that learns

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

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. Concisely, the main contributions of this work are as follows:

- We design a simple yet effective approach based on self-training, called `FedSTAR`, for exploiting large-scale unlabeled distributed data in a federated setting.
- We introduce adaptive confidence thresholding for generating pseudo-labels, which effectively reduces the number of inaccurate predicted annotations.
- We demonstrate through extensive evaluation that our technique is able to effectively learn generalizable audio models under a variety of label availability on diverse public datasets, namely Speech Commands [19], Ambient Context [20] and VoxForge [21].
- We show that our approach, `FedSTAR`, with as few as 3% labeled data, on average can improve recognition rate by 13.28% across datasets compared to the fully-supervised federated model.

2. METHODOLOGY

Problem Formulation: Formally, in semi-supervised FL setting, each of the K clients holds a labeled set, $\mathcal{D}_L^k = \{(x_{l_i}, y_i)\}_{i=1}^{N_{l,k}}$, where $N_{l,k}$ is the number of labeled data, x_{l_i} is an input instance, $y_i \,\epsilon\, \{1, \cdots, \mathcal{C}\}$ is the corresponding label, and $\mathcal{C}$ is the number of label categories for the $\mathcal{C}$-way multi-class classification problem. Besides, client holds a set of unlabeled samples denoted as $\mathcal{D}_U^k = \{x_{u_i}\}_{i=1}^{N_{u,k}}$, where, $N_{u,k}$ is the number of unlabeled data. Here, $N_k = N_{l,k} + N_{u,k}$ is the total number of data samples stored on the k-th client and $N_{l,k} \ll N_{u,k}$. Let $p_\theta(y \mid x)$ be a neural network that is parameterized by weights θ that predicts softmax outputs $\widehat{y}$ for a given input x. 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 is:

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

where $\mathcal{L}_{s_\theta}(\mathcal{D}_L^k)$ is the loss terms from supervised learning on the labeled data held by the k-th client, and $\mathcal{L}_{u_\theta}(\mathcal{D}_U^k)$ 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. For the FedAvg algorithm, parameter γ_k is equal to the ratio of client's local data N_k over all training samples $\left(\gamma_k = \frac{N_k}{N}\right)$.

Algorithm 1 `FedSTAR`: Federated Self-training for Audio Recognition. Here, l and u are equally sized on-device labeled and unlabeled batches respectively, β controls the contribution of unlabeled data during training, and η is the learning rate.

```
Server initialization of model $G$ with model weights $\theta_0^G$
for $i = 1, \dots, R$ do
    Randomly select $K$ clients to participate in round $i$
    for each client $k \in K$ in parallel do
        $\theta_i^k \leftarrow \theta_i^G$
        $\theta_{i+1}^k \leftarrow$ ClientUpdate($\theta_i^k$)
    end for
    $\theta_{i+1}^G \leftarrow \sum_{k=1}^{K} \frac{N_k}{N} \theta_{i+1}^k$
end for
procedure ClientUpdate($\theta$)
    for epoch $e = 1, 2, \dots, E$ do
        for batch $l \in \mathcal{D}_L$ and $u \in \mathcal{D}_U$ do
            $\widehat{y} \leftarrow \Phi\left(p_\theta(x_u), T\right)$
            $\theta \leftarrow \theta - \eta \nabla_\theta \left(\mathcal{L}_{CE}(y, p_\theta(y \mid x_l)) + \beta \cdot \mathcal{L}_{CE}(\widehat{y}, p_\theta(x_u))\right)$
        end for
    end for
end procedure
```

Federated Self-training (FedSTAR): We propose a self-training technique based on pseudo-labeling 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 recognition tasks, in order to learn from the labeled datasets $\mathcal{D}_L^k$ across all participating clients, we apply cross-entropy loss $\mathcal{L}_{s_\theta}(\mathcal{D}_L^k) = \mathcal{L}_{CE}\left(y, p_{\theta^k}(y \mid x_l)\right)$. Next, to learn from unlabeled data, we generate pseudo-labels $\widehat{y}$ for unlabeled data x_u of client k as:

$$\widehat{y} = \Phi(z, T) = \underset{i \in \{1, \dots, \mathcal{C}\}}{\arg\max} \left(\frac{e^{z_i/T}}{\sum_{j=1}^{\mathcal{C}} e^{z_j/T}} \right) \tag{2}$$

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. For the obtained pseudo-labels, we then perform standard cross-entropy minimization while using $\widehat{y}$ as targets, $\mathcal{L}_{u_\theta}(\mathcal{D}_U^k) = \mathcal{L}_{CE}\left(\widehat{y}, p_{\theta^k}(x_u)\right)$.

Concisely, in our `FedSTAR` algorithm, the clients' local update step is altered to learn from unlabeled datasets $\mathcal{D}_U^k$. 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 and the model's weights update is performed. The classical supervised categorical cross-entropy loss is minimized for batch l and afterward, pseudo-labels are produced using $\Phi(\cdot)$. With the creation of pseudo-labels $\widehat{y}$, the unlabeled batch u is then treated as a labeled batch $u^{'} = \{(u, \widehat{y})\}$, in which the client's model is further trained with standard cross-entropy loss. We

Table 1: Comparison of audio recognition models in centralized and federated settings. Average accuracy over three distinct trials is reported on test set.

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

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. 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.

Additionally, we propose an *adaptive confidence thresholding* method to diminish unsatisfactory performance due to training on faulty pseudo-labels. We use temperature scaling T to "*soften*" softmax output and generate confident predictions and employ an increasing confidence threshold τ to discard low-confidence pseudo-labels during training following a cosine schedule. Thus, we utilize solely high-confidence predictions when learning from unlabeled dataset D_U^k. Further details and an overview of our approach for the semi-supervised training procedure can be found in Algorithm 1.

3. EXPERIMENTS

Datasets: We use publicly available datasets to evaluate our method on a variety of audio recognition tasks. For all datasets, we use the suggested train/test split for comparability purposes. For ambient sound classification, we utilized the Ambient Acoustic Contexts dataset [20], in which sounds from ten distinct events are present. For keyword spotting task, we select the second version of Speech Commands dataset [19], where the model is trained to detect when a particular keyword is spoken out of a set of twelve target classes. Likewise, we use VoxForge [21] for the task of spoken language classification, which contains audio recordings in six languages - English, Spanish, French, German, Russian, and Italian.

Model Architecture: We use a convolutional neural network inspired by [22] with group normalization [23] 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 a prediction on an audio clip, we average over the predictions of non-overlapping segments of an entire audio. The model 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 in order to perform a joint 1×1 convolution. Using this scheme, the model can capture fine-grained features from each dimension and learn high-level features from their shared output. Furthermore, we apply L2 regularization with a rate of 0.0001 in each convolution 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 learning rate of 0.001 to optimize categorical cross-entropy.

To simulate a federated environment, we use the Flower framework [24] and utilize FedAvg [2] as an optimization algorithm to construct the global model from clients' local updates. We select a number of parameters to control the federated settings of our experiments. Those parameters are: 1) K - 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) β - influence of unlabeled data over training process, 8) T - temperature scaling parameter, and 9) τ - predictions confidence threshold. Across all `FedSTAR` experiments, we fixed $T = 4$, while we set τ to initialize from 0.5 and gradually increase to a maximum of 0.9 during training, following a cosine schedule.

Evaluation Strategy: In fully supervised federated experiments, where the entire dataset is available, the labeled instances are randomly distributed across the available clients. In experiments, where the creation of a labeled subset from the original is required (L <100%), we keep the dataset's initial class distribution ratio to avoid tempering with dataset characteristics. Likewise, in `FedSTAR` , the unlabeled subset consists of the dataset's remaining samples after extracting the labeled samples. In our experiments, both the labeled and unlabeled subsets are dispensed at random over the available clients. For a more rigorous evaluation, we manage any randomness during the data partitioning and training procedures by passing a specific seed, while 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 our results.

Results: We perform experiments in both centralized and fully-supervised federated settings to construct a high-quality *supervised* baseline. In federated settings, the FL parameters were set to R=100, q=80%, E=1 and σ=25%, while K varied, which is frequent in most real-life FL scenarios. Thus, we explore how the federated model behaves as clients progressively increase and the available local data become yet more distributed, affecting the performance of FL [25]. The resulting accuracy on the test set is presented in Table 1 and in the supervised federated row (L=100%) of Table 2 for centralized and federated experiments, respectively. Following, we evaluate `FedSTAR` to determine the obtained improvements versus

Table 2: Experiments on federated settings using `FedSTAR` with varying number of clients and label's availability. Average accuracy over three distinct trials on test set. Federated parameters are set to q=80%, σ=25%, β=0.5, E=1, R=100.

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**

Table 3: `FedSTAR` performance evaluation under varying class availability. The class distribution resembles a uniform distribution with mean μ=3 and variance σ_c. Average accuracy over three distinct trials is reported. Federated parameters are set to $\beta = 0.5$, R=100, K=15, q=80% and E=1.

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

a fully-supervised federated approach in non-i.i.d. settings. We run experiments, where L is varied from 3% up to 50%. To illustrate the performance gain of `FedSTAR` in comparison to the supervised FL regime, we performed experiments with identical labeled subsets,where, the unlabeled instances remained unexploited.

In Table 2, we observe that our method can utilize unlabeled audio data to improve the model's performance across all datasets significantly. We note that with L=5%, `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%). Comparing `FedSTAR` to the supervised FL for $L \leq 5\%$, we observe a notable improvement in accuracy across all cases. As we increase the percentage of labeled examples, the same trend is also present, although the performance gap between `FedSTAR` and supervised FL narrows considerably. In addition, while K increases and the labeled subset of each client shrinks, we notice that the `FedSTAR` models accuracy remains relatively unaffected. Consequently, `FedSTAR` can be applied in a label-scarce federated environment to boost the performance by learning from unlabeled data, independent of the audio recognition task.

We further investigate the performance of `FedSTAR` in highly non-i.i.d. setting, where the data distribution is skewed, both in terms of labels distribution and quantity of data per client. We perform experiments on the Speech Commands dataset with K=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 μ=3 and fluctuating variance σ_c from 0% to 50% as our class availability distribution across clients. For Speech Commands dataset, which holds 12 classes, choosing μ=3 results in clients obtaining only a few labeled samples from a restricted amount of classes (on average, three classes). This setup matches a real-world data distributions in pragmatic applications, where per-client labeled data from only few classes are available. From the results presented in Table 3, we note that `FedSTAR` can effectively exploit the available on-device unlabeled instances in a highly non-i.i.d. distributions. Comparing the `FedSTAR` with a fully supervised FL counterpart, we notice a substantial improvement in accuracy in all cases.

4. CONCLUSIONS

We propose a data-efficient federated self-training approach to learn audio recognition models with a few on-device labeled samples. Despite its simplicity, we demonstrate that our method is highly effective for semi-supervised learning on various tasks and we think it can be especially useful for applications, where the expertise is missing or the cost is high to annotate samples across devices like for instance in the medical domain.

5. REFERENCES

[1] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon, "Federated learning: Strategies for improving communication efficiency," 2017.

[2] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," 2017.

[3] David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau, "Federated learning for keyword spotting," 2019.

[4] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays, "Applied federated learning: Improving google keyboard query suggestions," 2018.

[5] Andrew Hard, Kanishka Rao, Rajiv Mathews, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage, "Federated learning for mobile keyboard prediction," *CoRR*, vol. abs/1811.03604, 2018.

[6] Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays, "Federated learning for emoji prediction in a mobile keyboard," *CoRR*, vol. abs/1906.04329, 2019.

[7] Andrew Hard, Kurt Partridge, Cameron Nguyen, Niranjan Subrahmanya, Aishanee Shah, Pai Zhu, Ignacio Lopez Moreno, and Rajiv Mathews, "Training keyword spotting models on non-iid data with federated learning," 2020.

[8] Hossein Hosseini, Sungrack Yun, Hyunsin Park, Christos Louizos, Joseph Soriaga, and Max Welling, "Federated learning of user authentication models," 2020.

[9] Dan Stowell, Yannis Stylianou, Mike Wood, Hanna Pamula, and Hervé Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *CoRR*, vol. abs/1807.05812, 2018.

[10] Justin Chan, Thomas Rea, Shyamnath Gollakota, and Jacob E Sunshine, "Contactless cardiac arrest detection using smart devices," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–8, 2019.

[11] Yilun Jin, Xiguang Wei, Yang Liu, and Qiang Yang, "Towards utilizing unlabeled data in federated learning: A survey and prospective," 2020.

[12] Xiaojin Zhu and Andrew Goldberg, *Introduction to Semi-Supervised Learning*, vol. 3, 01 2009.

[13] Romain Serizel, Nicolas Turpault, Hamid Eghbal-Zadeh, and Ankit Parag Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," *arXiv preprint arXiv:1807.10501*, 2018.

[14] Dong-Hyun Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, 2013, vol. 3.

[15] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," 2015.

[17] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang, "Federated semi-supervised learning with inter-client consistency," 2020.

[18] Zewei Long, Liwei Che, Yaqing Wang, Muchao Ye, Junyu Luo, Jinze Wu, Houping Xiao, and Fenglong Ma, "Fedsemi: An adaptive federated semi-supervised learning framework," 2020.

[19] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *CoRR*, vol. abs/1804.03209, 2018.

[20] Chunjong Park, Chulhong Min, Sourav Bhattacharya, and Fahim Kawsar, "Augmenting conversational agents with ambient acoustic contexts," in *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, New York, NY, USA, 2020, MobileHCI '20, Association for Computing Machinery.

[21] Ken MacLean, "Voxforge," *Ken MacLean.[Online]. Available: http://www.voxforge.org/home.[Acedido em 2012]*, 2018.

[22] Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, and Dominik Roblek, "Self-supervised audio representation learning for mobile devices," *arXiv preprint arXiv:1905.11796*, 2019.

[23] Yuxin Wu and Kaiming He, "Group normalization," 2018.

[24] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Titouan Parcollet, and Nicholas D Lane, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.

[25] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra, "Federated learning with non-iid data," 2018.