

Federated Self-training for Semi-supervised Audio Recognition

Vasileios Tsouvalas · Aaqib Saeed · Tanir Ozcelebi

Eindhoven University of Technology, The Netherlands

ACM TECS · Vol. 21, No. 6, Article 74 · 2022

01 Motivation

The problem

Scarce annotations leave most on-device audio unused by supervised federated learning (FL). The challenge is to learn from this unlabeled audio without collecting users’ recordings centrally.

Contributions

- **Joint local learning** uses labeled and unlabeled audio within one model.
- **Confidence filtering** adjusts pseudo-label selection as training progresses.
- **Label-budget evaluation** tests client distributions and optional public-audio pretraining.

Client data assumption

Each client holds few labeled clips and many unlabeled clips, assumed to share its local distribution. Training seeks one shared audio classifier.

Optional initialization

Public unlabeled audio can initialize the model through self-supervised pretraining. Experiments compare this optional initialization with random model weights.

Scope

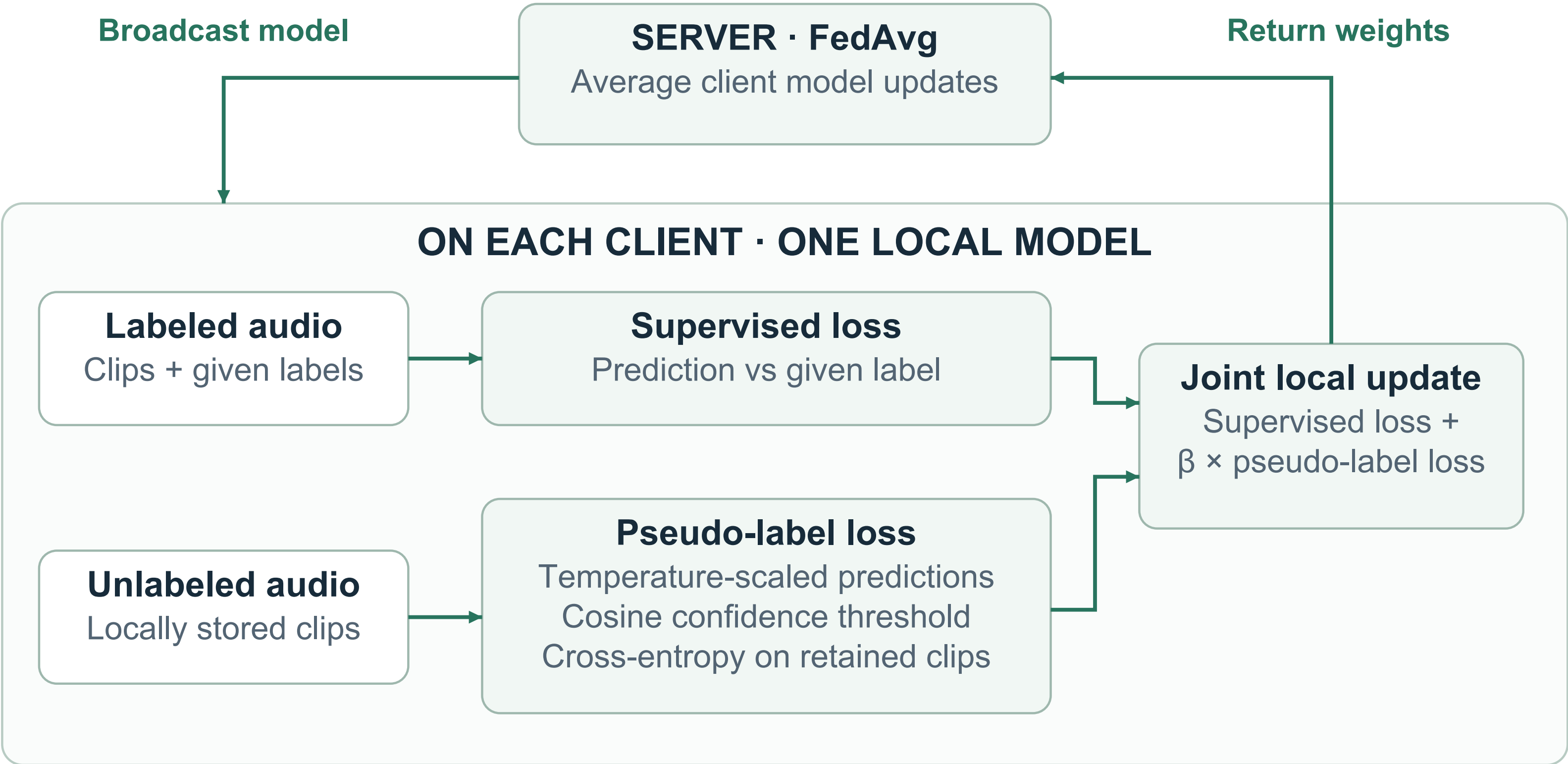
Training uses simulated clients. Class-skew tests keep unlabeled clips randomly distributed, so they do not establish robustness to matching labeled- and unlabeled-data skew.

Notation

cross-entropy (CE)

02 FedSTAR

One local model learns from true labels and pseudo-labels



Technical core

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

One client model uses true labels y and confident pseudo-labels $\hat{y}$. β weights the unlabeled-audio loss; both terms share one backward pass.



Paper



Code

03 Evaluation

3%

labeled data
15-client comparison

86.82%

Speech Commands accuracy
FedSTAR at 3% labels

Learning with 3% labels

Dataset	Supervised FL (%)	FedSTAR (%) (Ours)
Ambient Context	33.03 ± 0.002	49.54 ± 0.005
Speech Commands	62.98 ± 0.003	86.82 ± 0.006
VoxForge	54.26 ± 0.002	55.82 ± 0.011

3% labels, 15 clients, 100 rounds, 80% participation. Three-run mean accuracy; printed ± terms are described as variance.

Pretraining across label budgets

Labels	Random init. (%)	Pretrained (%)
3%	85.40 ± 0.008	85.17 ± 0.001
5%	88.52 ± 0.005	88.28 ± 0.002
20%	92.24 ± 0.012	93.67 ± 0.005
50%	92.26 ± 0.005	93.83 ± 0.003

Speech Commands; 50% of remaining unlabeled clips, 15 clients, 100 rounds. Three-run means; ± terms denote reported variance.

Takeaway

At 3% labels, **self-training improves all three tasks**, but gains vary widely. Public-data pretraining can help early convergence, while its **final-accuracy benefit depends on the data budget**.