

# Privacy-preserving Speech Emotion Recognition through Semi-Supervised Federated Learning

Vasileios Tsouvalas · Tanir Ozcelebi · Nirvana Meratnia

Eindhoven University of Technology

EmotionAware · IEEE PerCom Workshops 2022

## 01 Motivation

### The problem

Emotion annotation is costly, leaving much device speech unlabeled. Speech emotion recognition (SER) also involves sensitive recordings, motivating collaborative training that keeps raw audio on devices.

### Contributions

- **Attention-based features** capture spectral, temporal and channel cues from speech.
- **Device-specific confidence** filters pseudo-labels according to training progress and client participation.
- **Fixed-budget comparisons** test four speaker/client splits with only 10% labels.

### Local data, no server corpus

Each client has a small labeled subset and more unlabeled speech. The approach requires no server-side training data.

### Speaker variation

Random and per-speaker partitions test different client distributions. Seen- and unseen-speaker evaluation separates familiar speakers from speakers absent during training.

### Scope

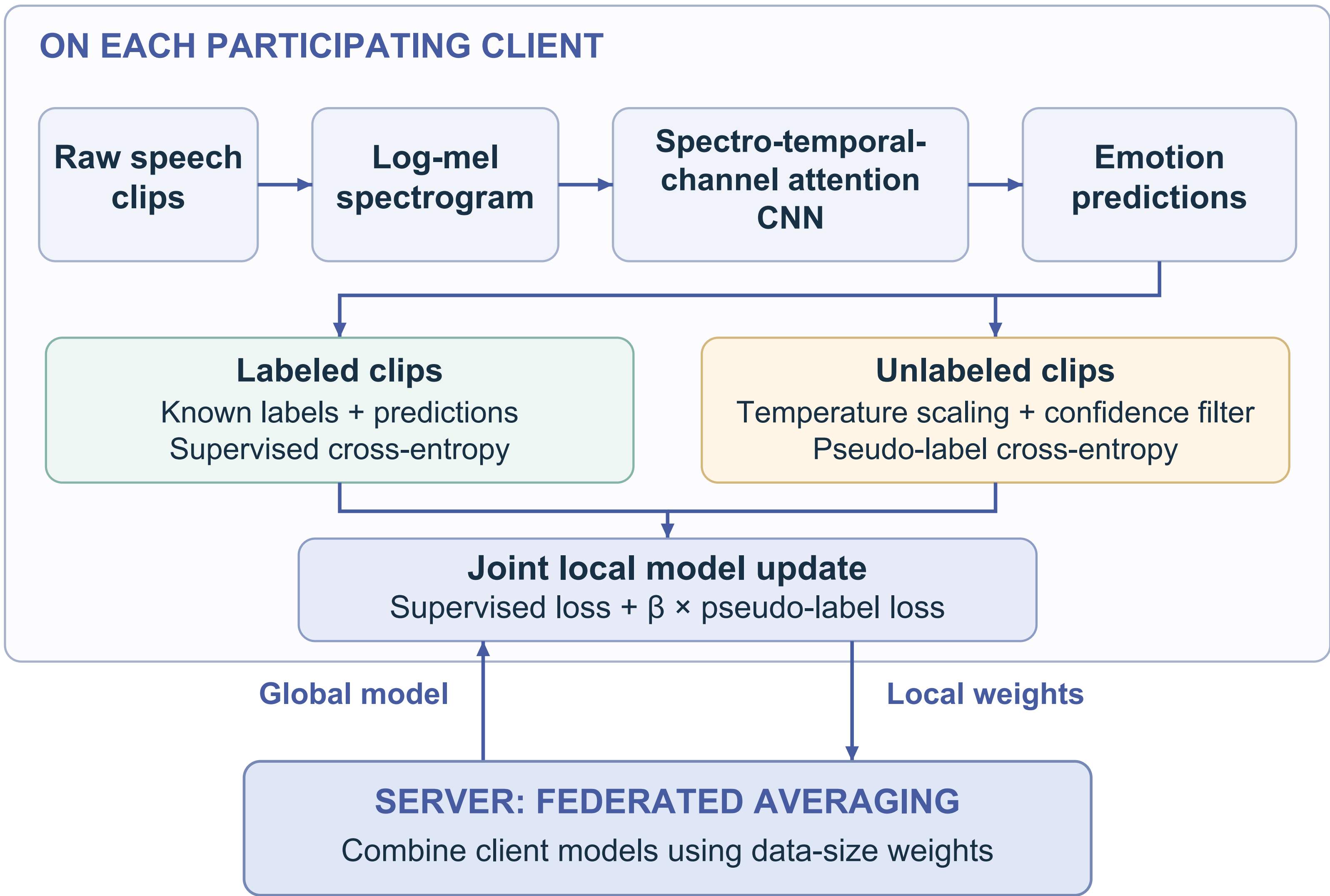
Evidence comes from one four-emotion corpus without reported dispersion. Keeping raw speech local does not provide differential privacy or protection against model-update attacks.

### Notation

convolutional neural network (CNN) · federated learning (FL) · cross-entropy (CE)

## 02 FedSER

### Use unlabeled speech without centralizing raw recordings



### Technical core

$$\mathcal{L}_k = \mathcal{L}_s(D_L^k) + \beta \mathcal{L}_u(D_U^k)$$

Client  $k$  combines supervised loss on labeled speech  $D_L^k$  with loss on confident pseudo-labels from  $D_U^k$ ;  $\beta$  sets their relative influence.



Paper



Code

## 03 Evaluation

10%

**labeled speech**

Remaining clips used unlabeled

56.83%

**unweighted accuracy**

Unseen speakers; per-speaker clients

### Unlabeled speech helps at a fixed label budget

| Speaker/client split  | Supervised 10% | Self-training 10% (Ours) | Supervised 100% |
|-----------------------|----------------|--------------------------|-----------------|
| Unseen / random       | 46.72          | <b>53.94</b>             | 60.52           |
| Unseen / per speaker  | 46.55          | <b>56.83</b>             | 59.55           |
| Overlap / random      | 51.58          | <b>59.04</b>             | 64.78           |
| Overlap / per speaker | 49.97          | <b>59.72</b>             | 62.79           |

IEMOCAP improvised speech, four emotions; five-run mean unweighted accuracy (%). Random/per-speaker partitions use 10/8 clients.

### Fully labeled reference models

| Same architecture | Unseen speakers (%) | Overlapping speakers (%) |
|-------------------|---------------------|--------------------------|
| Centralized       | 62.58               | 69.04                    |
| Federated         | 60.52               | 64.78                    |

100% labels, random partitions; mean unweighted accuracy. FL: 10 clients, 100 rounds, 80% participation, one local step.

### Takeaway

**Self-training improves every 10%-label comparison**, while remaining below the fully labeled reference. It keeps **raw recordings local**; the experiments establish no formal privacy guarantee.