2738

End-to-End Speech Recognition from Federated Acoustic Models

Yan Gao[1], Titouan Parcollet[2], Salah Zaiem[3], Javier Fernandez-Marques[4], Pedro PB Gusmao[1], Daniel J. Beutel[5], Nicholas D. Lane[1]

[1]University of Cambridge, [2]Université Avignon, [3]Telecom Paris, [4]University of Oxford, [5]Adap GmbH

TL;DR

- We quantitatively compare LibriSpeech to Common Voice towards a realistic FL setup to highlight the need for a shift in the evaluation of FL-based ASR models.
- The first study on attentional Seq2Seq E2E ASR model is conducted for FL scenarios. Concretely, we evaluate both *cross-silo* and *cross-device* FL with up to 4K clients on the naturally-partitioned and heterogeneous French and Italian subsets of Common Voice. This 4K client *cross-device* represents the largest scale FL ASR experiment of its kind ever performed.
- A first adapted aggregation strategy based on WER is proposed, integrating the specificity of ASR to FL.

ASR in FL: Background and Challenges

- FL is a form of distributed ML where nodes are edge devices such as smartphones, tablets or other IoT devices.
- On-device speech data is extremely non-IID by nature (e.g. different acoustic environments, words being spoken, microphones, etc.).
- SOTA E2E ASR models are computationally intensive and not suited for the on-device training phases of FL.
- E2E ASR training is difficult and very sensitive during early stages of optimisation due to the complexity of learning a proper alignment.
- These three traits make it very challenging to train ASR models completely from scratch. In fact, many works are evaluated on unrealistic datasets (w.r.t FL) such as LibriSpeech.

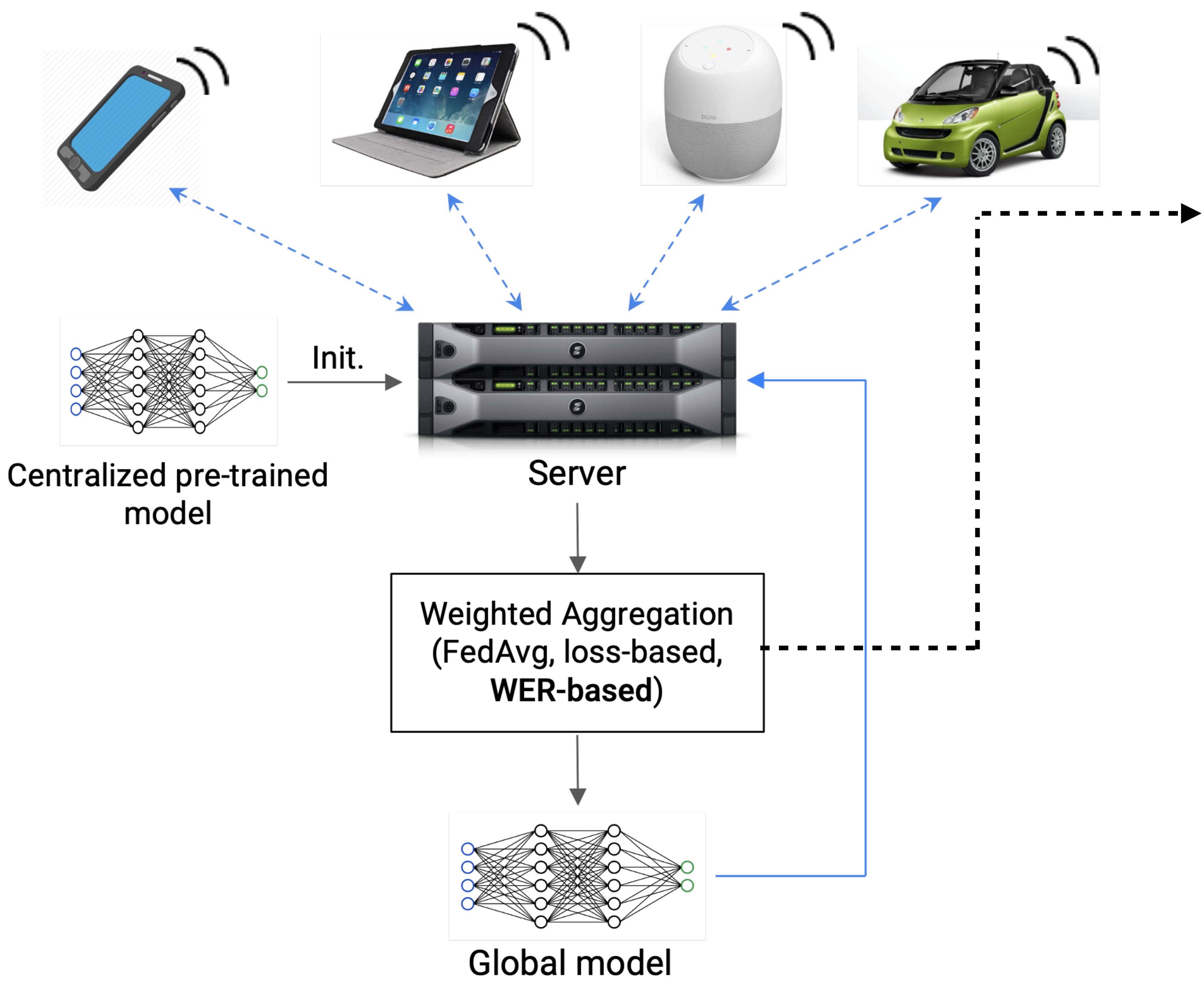
Building Realistic FL Environments

Dataset: Common Voice *French* (*Fr*) and Italian (*It*) set.

Centralised Pre-training: Train on *half* of the data samples with a small subset of speakers (*warm-up*). The other half data is for FL.

***Cross-silo* FL**: Splitting dataset in **10** random partitions with no overlapping speakers.

***Cross-device* FL**: 1) **Single** speaker using their individual devices. Dividing datasets based on users ID into **4095** and **649** partitions for *Fr* & *It*; 2) **Two** users per device. Splitting **2036** clients for *Fr*.



Init.

Centralized pre-trained model

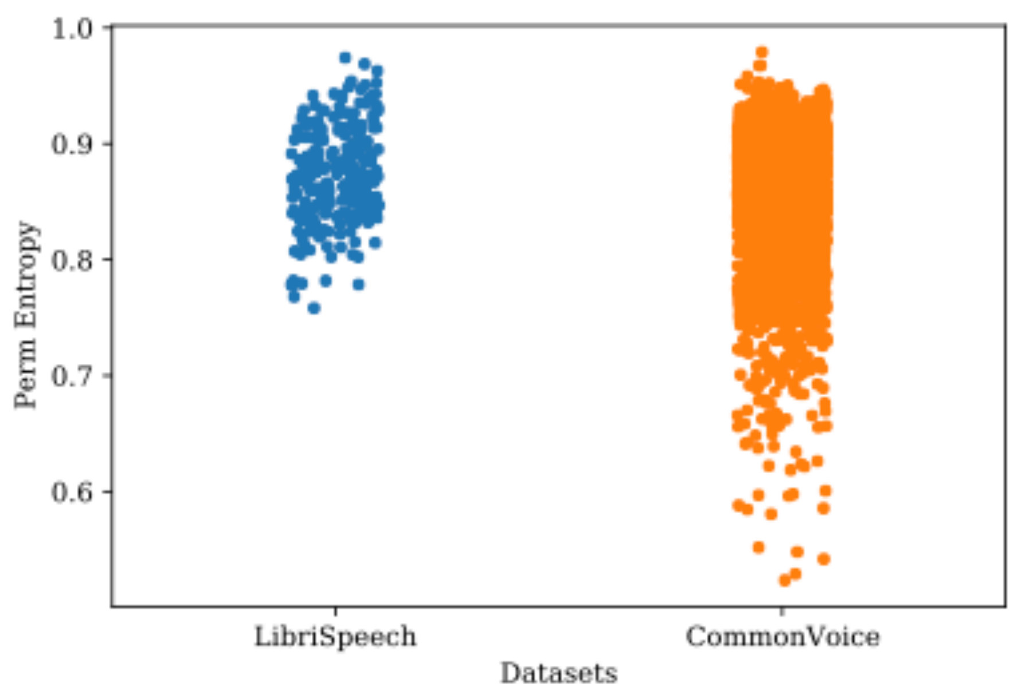
Server

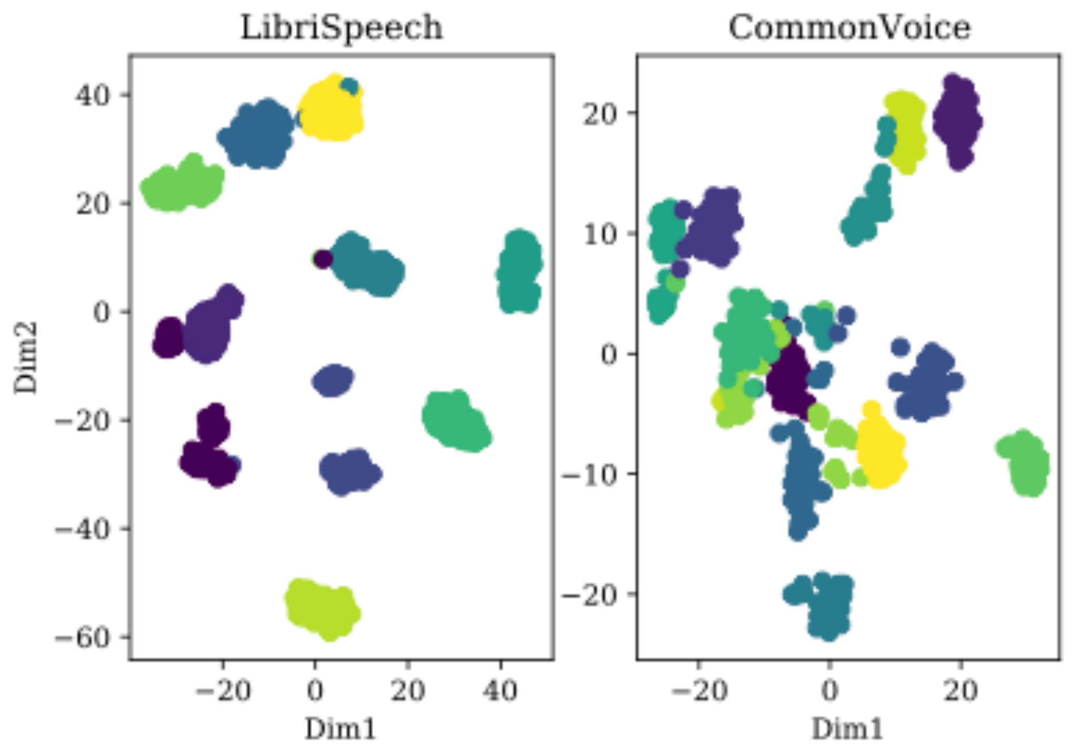
Weighted Aggregation (FedAvg, loss-based, WER-based)

Global model

Why Common Voice, not LibriSpeech?

- Most existing works use LibriSpeech (LS) for federated ASR model training, but we argue that Common Voice (CV) is closer to natural FL conditions than LS as much stronger variations are observed both intra- and inter-clients.
- CV shows a heavy-tailed distribution of permutation entropy values as a consequence of the bigger diversity of recording settings (left below).
- TSNE representation of embedded speech utterances via pre-trained speaker embeddings highlights the clustering difficulties in CV (right below).





Aggregation Weighting Strategies

1) FedAvg

$$\alpha^{(k)} = \frac{n_k}{\sum_{k=1}^{K} n_k}$$

2) Loss-based

$$\alpha^{(k)} = \frac{\exp\left(-\mathcal{L}_k\right)}{\sum_{k=1}^{K} \exp\left(-\mathcal{L}_k\right)}$$

3) WER-based

$$\alpha^{(k)} = \frac{\exp\left(1 - wer_k\right)}{\sum_{k=1}^{K} \exp\left(1 - wer_k\right)}$$

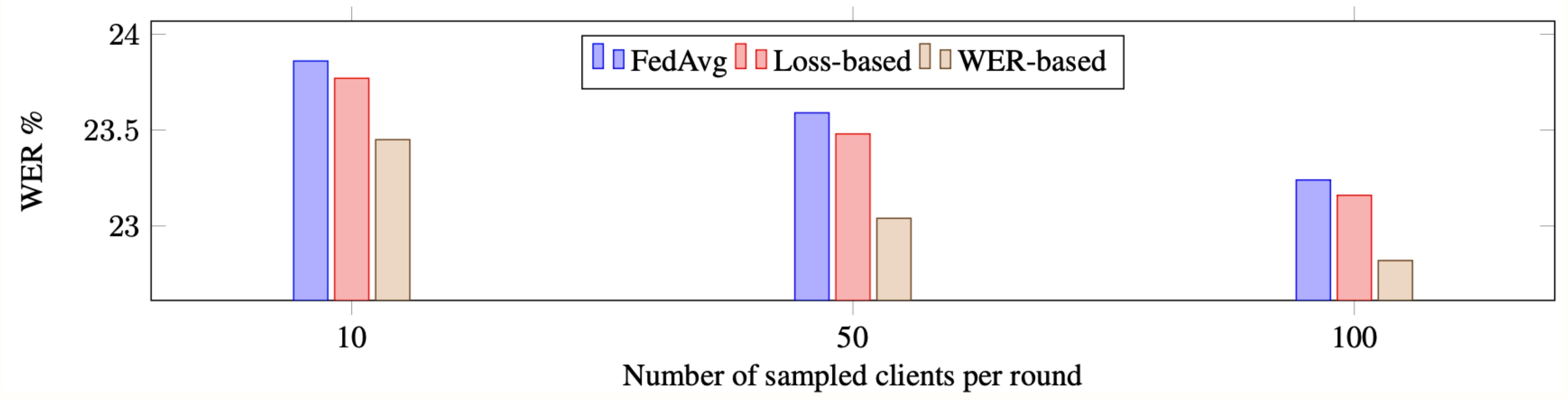
$\alpha^{(k)}$ represents the weighting for client k; n_k is the number of samples on client k; L_k is the averaged training loss from client k; wer_k is the validated WER on client k.

Experimental Results

- Training on the entire dataset in a centralised way gives us the lower bound WER.
- *Cross-silo* setting obtains better performance than *cross-device* setting due to its similar client distribution.
- WER-based methods achieve lower WER in all settings. This could be explained by the nature of the strategy which directly optimise the model toward the relevant metric for speech recognition.

Training Scenario		Fr WER(%)	It WER (%)
Centralized	All data (lower bound)	20.18	17.40
	1st half (*warm-up*)	25.26	25.90
	2nd half (*post warm-up*)	20.94	24.86
10-clients FL *Cross-silo*	Standard FedAvg	21.26	20.97
	Loss-based aggregation	21.10	20.86
	WER-based aggregation	**20.99**	**19.98**
2k-clients FL *Cross-device*	Standard FedAvg	22.83	—
	Loss-based aggregation	22.67	—
	WER-based aggregation	**22.42**	—
4k-clients FL *Cross-device*	Standard FedAvg	23.24	24.32
	Loss-based aggregation	23.16	24.23
	WER-based aggregation	**22.82**	**23.86**
From scratch	FedAvg, FedAdam	100	100

- We compare the impact of the number of selected clients per round on the most challenging setup, 4K clients in *French* set.
- *Higher values of selected clients produce better WER.*


FedAvg
Loss-based
WER-based
WER %
Number of sampled clients per round