# END-TO-END SPEECH RECOGNITION FROM FEDERATED ACOUSTIC MODELS

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

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

Deep neural networks are now widely adopted in state-of-the-art (SOTA) ASR systems [1]. This success mostly relies on the centralised training paradigm where data needs first to be gathered from one single dataset before it can be used for training [2, 3]. Such an approach has a few clear benefits including fast training and, the ability to sample data in any preferred way due to the complete data visibility. However, recent concerns around data privacy along with the proliferation of both powerful mobile devices and low latency communication (e.g. 5G), has caused distributed training paradigms such as federated learning (FL) to receive more attention.

In FL, training happens at the source and training data is never sent to a centralised server. In typical FL, clients receive a copy of the global model and train it separately using their own local data. This process generates a set of weight updates that are then sent to a server, where updates are aggregated. This process is repeated for several rounds [4]. Being able to harvest information from numerous mobile devices without collecting users' data makes federated ASR systems a feasible and attractive alternative to traditional centralised training [5], whilst offering new opportunities to advance ASR quality and robustness given the unprecedented amount of user data available on-device. For example, such data could be leveraged to better adapt the ASR model to the users' usage, or improve the robustness of models to realistic and low resources scenarios [6].

Despite the growing number of studies applying FL on speech-related tasks [7, 8], very few of these have investigated its use for E2E ASR. Properly training E2E ASR models in a realistic FL setting comes with numerous challenges. First, it is notoriously complicated to train a deep model with FL on non independent and identically distributed data (non-IID) [6] and on-device speech data is extremely non-IID by nature (e.g. different acoustic environments, words being spoken, microphones, etc.). Second, SOTA E2E ASR models are computationally intensive and not suited for the on-device training phases of FL. Indeed, SOTA ASR systems rely on large Transformers, Transducers or sequence-to-sequence (Seq2Seq) models. Finally, 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 [9]. To our best knowledge, existing works typically approach these challenges by relaxing one or more of these challenges in their experimental design. In fact, many works [10, 11] are evaluated on unrealistic datasets (*w.r.t* FL) such as LibriSpeech (LS), which only contain speakers reading books without background noise or other characteristics typical of FL settings.

In this work, we highlight the need for researchers to move away from clean speech corpora when evaluating FL-based ASR systems. We perform the first quantitative comparison of LS against a new alternative: the Common Voice (CV) dataset, which provides a large, heterogeneous and uncontrolled set of speakers who used their own devices to record a set of sentences; naturally fitting to FL with various users, acoustic conditions, microphones and accents. We discover, that under realistic FL conditions captured by the CV dataset, conventional FL aggregation (used in prior work like [10, 11]) struggle to even converge during training. In response, we devise a novel ASR system able to cope with such realistic FL conditions. We show our approach works under both a *cross-silo* and a *cross-device* (i.e. large number of clients with few naturally non-IID data) FL setting while training a SOTA E2E ASR system.

## 1. REFERENCES

[1] Akshi Kumar et al, "A survey of deep learning techniques in speech recognition," in *ICACCCN 2018*. IEEE, 2018, pp. 179–185.

[2] Awni Hannun et al, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[3] Dario Amodei et al, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[4] Brendan McMahan et al, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[5] Jakub Konečný et al, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[6] Peter Kairouz et al, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[7] Andrew Hard et al, "Training keyword spotting models on non-iid data with federated learning," *arXiv preprint arXiv:2005.10406*, 2020.

[8] David Leroy et al, "Federated learning for keyword spotting," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6341–6345.

[9] Andrew Rosenberg et al, "End-to-end speech recognition and keyword search on low-resource languages," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5280–5284.

[10] Dhruv Guliani et al, "Training speech recognition models with federated learning: A quality/cost framework," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3080–3084.

[11] Dimitrios Dimitriadis et al, "A federated approach in training acoustic models," in *Proc. Interspeech*, 2020.