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

ABSTRACT

Training Automatic Speech Recognition (ASR) models under federated learning (FL) settings has attracted a lot of attention recently. However, the FL scenarios often presented in the literature are artificial and fail to capture the complexity of real FL systems. In this paper, we construct a challenging and realistic ASR federated experimental setup consisting of clients with heterogeneous data distributions using the French and Italian sets of the CommonVoice dataset, a large heterogeneous dataset containing thousands of different speakers, acoustic environments and noises. We present the first empirical study on an attention-based sequence-to-sequence End-to-End (E2E) ASR model with three aggregation weighting strategies – standard FedAvg, loss-based aggregation and a novel word error rate (WER)-based aggregation, compared in two realistic FL scenarios: *cross-silo* with 10 clients and *cross-device* with 2K and 4K clients. This 4K *cross-device* ASR experiment is the largest ever performed. Our first-of-its-kind analysis on E2E ASR from heterogeneous and realistic federated acoustic models provides the foundations for future research and development of realistic FL ASR applications.

Index Terms— End-to-end ASR, federated learning

1. INTRODUCTION

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, 4]. 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 [5, 6, 7]. 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 [6], 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 [7].

Despite the growing number of studies applying FL on speech-related tasks [8, 9, 10, 11, 12, 13, 14], 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) [7, 15, 16] 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 [17, 18], Transducers [19, 20] or sequence-to-sequence (Seq2Seq) models [21, 22]. 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 [23, 24]. 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 [14, 10] are evaluated on unrealistic datasets (*w.r.t* FL) such as LibriSpeech (LS) [25], 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 [26], 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 [14, 10]) 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.

Our contributions are as follows. First, we quantitatively compare LS to CV towards a realistic FL setup to highlight the need for a shift in the evaluation of FL-based ASR models. Second, 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 CV. This 4K client *cross-device* represents the largest scale FL ASR experiment of its kind ever performed. Third, a first adapted aggregation strategy based on WER is proposed, integrating the specificity of ASR to FL. Finally, we release the code using Flower [27] and SpeechBrain [28] to facilitate replication and future research[1].

2. FEDERATED TRAINING OF ACOUSTIC MODELS

The process of training an E2E acoustic model using federated learning follows four steps: 1) Following [10], model weights are ini-

[1]github.com/yan-gao-GY/Flower-SpeechBrain

tialised with a pre-training phase on a centralised dataset; 2) The centralised server samples K clients from a pool of M clients and uploads to them the weights of the model. 3) The clients train the model for t_{local} local epochs in parallel based on their local user data and send back the new weights to the server. 4) The server aggregates the weights and restart at step 2. This procedure is executed for T rounds until the model converges on a dedicated validation set (e.g. local to each client or centralised).

2.1. Federated Optimisation

For each training round, each client $k \in K$, containing n_k audio samples, runs $t \in [0, t_{local}]$ iterations with learning rate η_l to locally update the model,

$$w_{t+1}^{(k)} = w_t^{(k)} - \eta_l \tilde{g_k}, \tag{1}$$

with w_k the local model weights of client k, and $\tilde{g_k}$ the average gradient over local samples. After training for t_{local} local epochs in the global round T, the updated weights $w_T^{(k)}$ of the client k are sent back to the server. Then, the local gradient $g_T^{(k)}$ is computed as:

$$g_T^{(k)} = w_T^{(k)} - w_{T-1}. \tag{2}$$

Then, the gradients from all clients are aggregated as follows:

$$\Delta_T = \sum_{k=1}^{K} \alpha_T^{(k)} g_T^{(k)}, \tag{3}$$

where $\alpha_T^{(k)}$ denotes different weighting strategies described in Section 2.2. The updated global model weights $w_T = w_{T-1} - \eta_s \Delta_T$, are computed with a server learning rate η_s.

During ASR FL training, especially with heterogeneous data, the global model may deviates away from the original task or simply does not converge [7, 15, 16], and therefore lead to performance degradation. To alleviate this issue, and motivated by [10], we propose an additional training iteration over a small batch of held-out data on the server, after the standard model update procedure.

2.2. Weighting Strategies

Federated Averaging (FedAvg) [5] is a popular [29, 30, 31, 8, 9, 14] aggregation strategy by which model updates from each client are weighted by $\alpha_T^{(k)}$, the ratio of data samples in each client over the total samples utilized in the round:

$$\alpha_T^{(k)} = \frac{n_k}{\sum_{k=1}^{K} n_k}, \tag{4}$$

In realistic FL settings with heterogeneous client data distribution, some clients may contain skewed data not representing the global data distribution (e.g. audio samples with different languages or multiple speakers). As a result, the aggregated model might simply not converge if such clients have proportionally more training samples than others. For instance, in our experiments, all attempts to train an ASR system from scratch failed due to this issue requiring a prior pre-training phase of the acoustic model. Second, clients containing low quality data would introduce unexpected noise into the training process (e.g. extreme noise in the background). Either scenario could lead to model deviation in the aggregation step, which can not be solved via the standard FedAvg weighting method (Eq. 4). A potential solution, instead, is to use the averaged training loss as a weighting coefficient, thus reflecting the quality of the locally trained model. Intuitively, higher loss would indicate that the global model struggles to learn from the client's local data. More precisely, we compute the weighting with the *Softmax* distribution obtained from the training loss. Eq. 4 is modified as follows:

$$\alpha_T^{(k)} = \frac{\exp\left(-\mathcal{L}_k\right)}{\sum_{k=1}^{K} \exp\left(-\mathcal{L}_k\right)}. \tag{5}$$

The softmax simply ensures that we are giving a higher priority to well-performing clients. In the context of ASR, WER is commonly used as the final evaluation metric for the model instead of the training loss. We therefore propose a WER-based weighting strategy for aggregation. This approach utilizes the values $(1 - wer)$ obtained on the validation set as weighting coefficients $\alpha_T^{(k)}$:

$$\alpha_T^{(k)} = \frac{\exp\left(1 - wer_k\right)}{\sum_{k=1}^{K} \exp\left(1 - wer_k\right)}. \tag{6}$$

To ensure a fair weighting between clients suffering from varying utterance lengths, wer_k is computed every round on a unique and centralised set. This also prevents the aggregation noise induced by incorrect local transcriptions coming from the FL scenario. In this way, all clients are weighted w.r.t the same conditions.

3. COMMON VOICE AS A REALISTIC FL SETUP

In this section we first present the CV dataset used for the FL setup. Then, we quantitatively demonstrate that CV is a much more suitable corpus to advance FL research with its non-IID property than LS, motivating the need for a shift in the standard evaluation process.

3.1. Common Voice dataset

Both the French and Italian subsets of CV dataset (version 6.1) [26] are considered. Utterances are obtained from volunteers recording sentences all around the world, and in different languages, from smartphones, computers, tablets, etc. The French set contains a total of 328K utterances (475 hours) with diverse accents which were recorded by more than 10K French-speaking participants. The train set consists of 4212 speakers (425.5 hours of speech), while both valid and test sets contain around 24 hours of speech from 2415 and 4247 speakers respectively. The Italian set, on the other hand, is relatively small, containing 89, 21 and 22 hours of Italian training (748 speakers), valid (1219 speakers) and test (3404 speakers) data.

3.2. Setup analysis and LibriSpeech comparison

We argue that CV is closer to natural FL conditions than LS as much stronger variations are observed both intra- and inter-clients. While CV is a crowd-sourced dataset containing thousands of different acoustic conditions, microphones and noises, LS is a heavily controlled studio-quality corpus. The latter has been used by most research on FL ASR. We compare both datasets at three levels:

Low-level signal features. The selected features should be more descriptive of the background and recording conditions than speaker identity, as this is investigated when analysing clustering purity. Hence, we will consider: Loudness as it is highly linked to the microphone and the recording distance; the log of the Harmonicity to Noise Ratio (logHNR) as a proxy indicator of background noise; Permutation Entropy (PE) as it has been successfully used for microphone identification purposes[32].

The mean value of the signal feature is computed for every utterance by averaging the per-frame values. Then, for every client we compute the mean value and standard deviation (SD). The former distribution describes the inter-client variation while the latter describes the intra-client one. For the three considered features, the SD of the mean value per client distribution is higher for CV than for

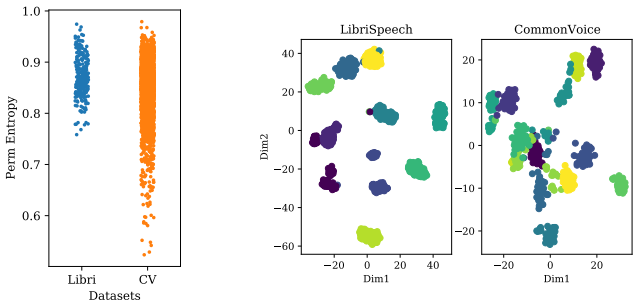

Fig. 1: *(Left)* Strip plot of the permutation entropy mean values per client in LS and CV. CV shows a heavy-tailed distribution as a consequence of the bigger diversity of recording settings. *(Right)* TSNE representation of embedded speech utterances. The colors correspond to the true clients (*i.e.* speakers).

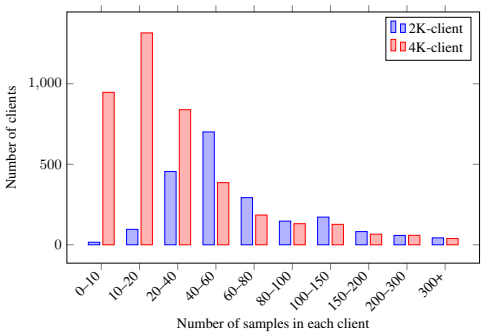

Fig. 2: Illustration of the sample distribution across the 2K-client and 4K-client FL settings from the French CV dataset.

LS, reaching 0.034, 11.466 and 0.053 for, respectively, Loudness, logHNR and PE on CV compared to 0.017, 9.096 and 0.040 on LS. Concerning the intra-client variation, the SD of the SD per client distribution is also higher for CV than for LS reaching 0.009, 2.69 and 0.014 against 0.007, 2.31 and 0.007 respectively for loudness, logHNR and PE. It is also interesting to note the heavy tailed distribution obtained with the PE for CV, as depicted in Fig. 1. Indeed, the Kurtosis reaches 4.16 on CV versus -0.13 for LS. In practice, this mean that many clients may be outliers for CV, drastically impacting FL with conventional aggregation mechanisms.

Blind Signal-to-Noise ratios. We further inspect the noise difference between two datasets through computing a blind Signal to Noise Ratio estimation. First, a 10th-order LPC approximation is computed for every sample. Second, the voiced chunks are detected using the Probabilistic YIN algorithm for F0 estimation [33]. Finally, considering only the voiced chunks that are simpler to approach with an LPC estimate, the noise in the blind SNR estimation is defined as the difference between the real signal and the LPC approximation. Following the trend observed with the signal features, CV shows a higher SD for the SNR mean values 18.47 compared to 10.32 for LS. Then, a bigger variation within recordings of the same client is observed. Indeed, the SD of the SDs obtained for each audio sample of the same client is higher in CV than in LS, with 6.54 compared to 3.82. This suggests a higher variability in the recording conditions with respect to the same client. Common Voice speakers may contribute from different places and devices.

Clustering purity. We compare the overlap of speakers via pre-trained speaker embeddings. For both datasets, speaker embeddings are computed using the *Tristounet* model [34, 35]. It is important to mention that *Tristounet* is not trained on LS or CV or audio book data. The embeddings are then clustered by the K-means algorithm with the number of centroids equal to the number of clients. The purity of the clusters is defined as the proportion of points that belong to the same client as the majority of its computed cluster. Purity reaches 0.77 on LS and 0.62 on CV. Fig. 1 shows a TSNE representation of the embeddings, and highlights the clustering difficulties in CV. This indicates that CV speakers are harder to separate. This confirms the two prior experiments using low-level audio features, as it suggests that varying signal features and recording conditions pollute the speech which leads to harder speaker identification.

The analysis in this section evidences the drastic differences in corpora between LS and CV. The latter better captures the complexity that FL systems would face when deployed in the real world.

4. EXPERIMENTAL SETTINGS

This section first present the architecture of the E2E attention-based Seq2Seq speech recognizer. Then, it describes the experimental setup of the FL environment alongside with key hyper-parameters.

4.1. E2E Speech Recognizer

The experiments are based on a Seq2Seq model trained with the joint connectionist temporal classification (CTC)-attention objective [21]. A typical ASR Seq2Seq model includes three modules: encoder,decoder and attention module. The encoder follows the CRDNN architecture first described [28] (*i.e. 2D CNN — LSTM — DNN*). The decoder is a location aware GRU with a single hidden layer. The full set of parameters describing the model are given in the GitHub repository. Models are jointly trained with CTC and cross entropy (CE) loss to predict sub-words units. No language model fusion is performed to properly assess the impact of the training procedure on the acoustic models.

4.2. Realistic Federated Learning

Based on the natural partitioning of the CV dataset we conduct two sets of experiments reflecting real usages of FL:

***Cross-silo* FL**. In this setting, clients are generally few, with high availability during all rounds and, often have similar data distribution [7]. Shared data is often independent and identically distributed. Our implementation follows that of [10], the dataset is split in 10 random partitions (*i.e.* one per client) with no overlapping speakers each containing roughly the same amount of speech data.

***Cross-device* FL**. This setup often involves thousands of clients having very different data distributions (non-IID), each participating in just a few rounds [7]. Here, we define two settings: First, we simulate a scenario of single speaker using their individual devices. To reproduce this, we naturally divide the training sets based on users ID into 4095 and 649 partitions for French and Italian, respectively. The second scenario allocates two users per device (e.g as in personal assistants or smart cars). For CV French, this lowers the number of clients to 2036. As depicted in Fig. 2, each setting drastically change the distribution of *low-resource* clients. The $4K$ setup offers a challenging scenario as most clients only contain very few samples.

4.3. Federated Learning for ASR: a hybrid approach

Training E2E ASR models in a FL setting is challenging, commonly requiring large datasets. Therefore, and as we experienced during our analysis, it is nearly impossible to train an E2E ASR model from scratch in a realistic FL setup. Table 1 shows that all the tested existing FL aggregation methods fail to converge without pre-training.

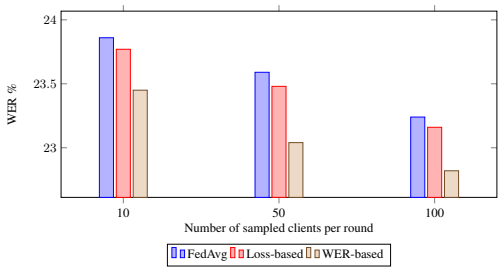

Fig. 3: Speech recognition performance when varying the number of sampled clients per round for the 4K-client setting on French CV.

This is due to the fact that most of the clients only contain few speech data, resulting in an extremely noisy gradient to learn the alignment from. To overcome this issue we first pre-train the model on half of the data samples. We do this by partitioning the original dataset into a small subset of speakers (with many samples) for centralised training (referred to subsequently as the *warm-up* dataset) and a larger subset of speakers (having fewer samples each) for the FL experiment. For CV French, the small subset contains 117 speakers, leaving the remaining 4095 speakers for FL. Such statistics are reduced down to 99 and 649 speakers for Italian. We argue that this scenario remains realistic as, in practice, centralised data is often available and can therefore be used to bootstrap the alignment.

The number of clients participating in each round influences the outcome of the experiments. To quantify this variation, we propose to vary the selected number of clients per round K from 10 to 100 for all weighting strategies on the $4K$ set. Then, we simply fix K with respect to the best WER obtained (*i.e.* 100) for the others setups. For the *cross-silo* setting, all clients are selected each round ($K = 10$). Defining the number of local epoch (i.e. on each client) for the FL experiment is not a non-trivial task [5]. In practice, we found that higher local epochs leads to clients over-fitting their own local data. Hence, clients are locally trained for only 5 epochs.

Training concurrently a large number of clients might become challenging. While models may be trained with CPUs or modest GPUs on real embedded hardware (e.g. RapsberryPi or NVIDIA Jetson), our FL setup allows us to run these workloads on modern GPUs running multiple clients concurrently on a single GPU with Flower [27] and SpeechBrain [28].

5. SPEECH RECOGNITION RESULTS

First, we compare the impact of the number of selected clients K per round on the most challenging setup (4K clients in French, the largest scale FL ASR experiment of its kind ever performed, Fig. 3). Conversely to the literature, higher values of K produce better WER. This is explained by the heterogeneity of the CV dataset, for which extremely noisy clients may perturb the averaging process with few clients per round. For the remaining of the experiments, $K = 100$.

Table 1 reports the results obtained across the different training setups. We notice that training on the entire dataset in a *centralised* way gives us the best WER with 20.18% and 17.40% for the French and Italian sets respectively, which is comparable to the current best literature [28]. This lower-bound is expected as the system has full visibility of the data and can sample the inputs in an almost IID fashion. On the other hand, when using only the *warm-up* dataset, we notice the effect of having fewer data points for training as the WER increases to 25.26% for the French set and 25.90% for the Italian set. This is expected as well as the system has now less data to learn from. This sheds some light on the inherent lower-bound limitations

Table 1: Speech recognition results on the centralised test sets of French (*"Fr"*) and Italian (*"It"*) CV datasets. *"User-based"* FL represents 4K clients for French and 649 for Italian. We find (see last row) that conventional FL methods fail to converge under CV.

Training Scenario		Fr WER (%) ↓	It WER (%) ↓
Centralised	All data (lower bound)	20.18	17.40
	1^{st} half (*warm-up*)	25.26	25.90
	2^{nd} half (post *warm-up*)	20.94	24.86
10-clients FL *Cross-silo*	FedAvg	21.26	20.97
	Loss-based	21.10	20.86
	WER-based	**20.99**	**19.98**
2K-clients FL *Cross-device*	FedAvg	22.83	—
	Loss-based	22.67	—
	WER-based	**22.42**	—
User-based FL *Cross-device*	FedAvg	23.24	24.32
	Loss-based	23.16	24.23
	WER-based	**22.82**	**23.86**
From scratch	FedAvg, FedAdam [36, 5]	100+	100+

of FL, limited to partial data observations in each round. The third centralised scenario trains the warmed-up model on the 2nd half of data in an on-line training fashion. This model provides a slightly lower WER compared to all FL models in French set. However, we should note that this is an unrealistic setting as training models in a centralised way would void all the privacy guarantees that FL offers. In particular, this model only gains 0.14% improvement in Italian set compared with the *warm-up* model. This indicates the difficulty of training model on the 2nd half data even in centralised fashion. The results on all FL settings exceed centralised training thanks to the centralised fine-tuning in between each round on the server side.

The effect of data visibility can indeed be seen in both *cross-silo* and *cross-device* scenarios, which do not have uniform access to data. However, since this problem is less severe in the former setup, we are still able to obtain a WER of 20.99% with the French set, which is very close to the centralised lower bound of 20.18%. The more challenging Italian set, on the other hand, obtains 19.98% WER with a 2.58% difference to the lower bound. As for the *cross-device* scenario, the effect of non-IID data distribution among devices leads to its best WER on French set being 22.43% and 22.82% in the 2K and 4K clients settings, even worse (23.86%) with the Italian set. We decided to not experiment ourselves with LS as prior work already did it [10] and has demonstrated that, in contrast to our findings with CV, FL experiments can even exceed the performance of centralised training. This indicates that the data distribution of LS is too simple to represent a realistic evaluation setup for FL.

Compared to different weighting strategies, WER-based and loss-based methods obtain a better performance in all settings, which indicates that weakening the effects of low-quality clients can assist the aggregation process in federated training with heterogeneous data distribution. Herein, we have two types of indicators reflecting the quality of clients. Tab. 1 shows that WER-based strategy (in bold) obtain the lowest WER in all settings. This could be easily explained by the nature of the strategy which directly optimise the model toward the relevant metric for speech recognition.

6. CONCLUSION

In this paper, we presented the first study for realistic FL scenarios on attention-based Seq2Seq E2E ASR models with three aggregation weighting strategies. We quantitatively compared LibriSpeech and Common Voice towards a realistic FL setup. All methods were evaluated with *cross-silo* and *cross-device* FL on two languages. Our work sets the foundations for future research of realistic FL ASR applications with an open source environment.

7. REFERENCES

[1] Akshi Kumar et al, "A survey of deep learning techniques in speech recognition," in *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. 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] Hagen Soltau et al, "Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition," *arXiv preprint arXiv:1610.09975*, 2016.

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

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

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

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

[9] 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.

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

[11] Xiaodong et al Cui, "Federated acoustic modeling for automatic speech recognition," *arXiv preprint arXiv:2102.04429*, 2021.

[12] Filip Granqvist et al, "Improving on-device speaker verification using federated learning with privacy," *arXiv preprint arXiv:2008.02651*, 2020.

[13] Zhenhou Hong et al, "Federated learning with dynamic transformer for text to speech," *arXiv preprint arXiv:2107.08795*, 2021.

[14] 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.

[15] Yue Zhao et al, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[16] Felix Sattler et al, "Robust and communication-efficient federated learning from non-iid data," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 9, pp. 3400–3413, 2019.

[17] Abdelrahman Mohamed et al, "Transformers with convolutional context for asr," *arXiv preprint arXiv:1904.11660*, 2019.

[18] Albert Zeyer et al, "A comparison of transformer and lstm encoder decoder models for asr," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 8–15.

[19] Mehryar Mohri et al, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.

[20] Eric Battenberg et al, "Exploring neural transducers for end-to-end speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 206–213.

[21] Suyoun Kim et al, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[22] Chung-Cheng Chiu et al, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[23] 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.

[24] Sameer Bansal et al, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 58–68.

[25] V. Panayotov et al, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[26] Rosana Ardila et al, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[27] Daniel J Beutel et al, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.

[28] Mirco Ravanelli et al, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[29] Tian Li et al, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[30] Samuel Horvath et al, "Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout," 2021.

[31] Xinchi Qiu et al, "A first look into the carbon footprint of federated learning," 2021.

[32] Gianmarco Baldini et al, "An Evaluation of Entropy Measures for Microphone Identification," *Entropy*, vol. 22, no. 11, 2020.

[33] Matthias Mauch et al, "Pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.

[34] Hervé Bredin, "Tristounet: Triplet loss for speaker turn embedding," 2017.

[35] Hervé Bredin et al, "pyannote.audio: neural building blocks for speaker diarization," 2019.

[36] Sashank Reddi et al, "Adaptive federated optimization," *arXiv preprint arXiv:2003.00295*, 2020.