Federated Learning for Speech Recognition: Revisiting Current Trends Towards Large-Scale ASR

Sheikh Shams Azam* **Martin Pelikan*** **Vitaly Feldman** **Kunal Talwar**

Jan "Honza" Silovsky **Tatiana Likhomanenko***

Apple

{s_azam,mpelikan,vitalyf,ktalwar,jsilovsky,antares}@apple.com

Abstract

While automatic speech recognition (ASR) has witnessed remarkable achievements in recent years, it has not garnered a widespread focus within the federated learning (FL) and differential privacy (DP) communities. Meanwhile, ASR is also a well-suited benchmark for FL and DP as there is (i) a natural data split across users by using speaker information; (ii) heterogeneous data across speakers close to practical settings; (iii) interplay between acoustic and language modeling; (iv) and it is a sequence-to-sequence task. Recent production-ready state-of-the-art models in ASR include *large* conformer and transformer models, the optimization of which is known to pose challenges even for central training. While the main trends and benchmarks in FL and DP focus on *small* models, we show the necessity of disentangling optimization and model size: the behavior of FL and DP for *large* models is different from the one for *small* models. We speculate that FL and DP are harder for *small* models due to harder optimization problems even in central training. In this paper, we analyze the key FL parameters (optimizers, training from scratch or a seed model pre-trained centrally, cohort size, data heterogeneity) and propose *first* benchmark of *FL with DP* in the context of *large* models in ASR. We examine the applicability of prior results and present an overview of observed departures from the trends in prior works and from training different ASR models. Through this work, we provide researchers and practitioners in the fields of FL and DP with valuable insights into the fundamental differences that may arise when applying FL and DP research to large-scale ASR training.

1 Introduction

Automatic speech recognition (ASR) with end-to-end (E2E) models made a remarkable achievements in recent years [1; 2; 3; 4; 5; 6]. Moreover, recent works incorporate ASR as a benchmark alongside vision and natural language processing (NLP) [7; 8; 9; 10]. However, ASR is not a widespread benchmark in federated learning (FL) [11] and differential privacy (DP) [12; 13; 14] communities. Scarce prior works applied FL to E2E ASR [15; 16; 17; 18; 19; 20]; most of these studies pointed out that E2E ASR is challenging for FL especially due to the inherently heterogeneous data. Besides, we are not aware of any studies on DP or FL with DP for ASR, although audio data reveal information

*Equal contribution.

International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023 (FL@FM-NeurIPS'23). This workshop does not have official proceedings and this paper is non-archival.

not only about the content but also could be used to derive other pieces of sensitive information such as biometric identity, emotions, health condition and others [21].

Similar to [18], we argue that ASR is also a well-suited benchmark for FL and FL with DP as there is (i) a natural split across users by using speaker information available for most public datasets; (ii) high data heterogeneity (see Figure 1 and discussion in [18]), covering variation in acoustic features across speakers, acoustic environments, recording conditions, and amount of data; (iii) interplay between acoustic and language modeling; (iv) unique challenges of ASR in the context of DP that we show in this paper and that can advance FL with DP more broadly. Properties (i)-(iii) stand out for ASR public datasets compared to others used by FL and DP communities in publications. For example, in computer vision, we often do not have a natural data split across users for standard datasets such as MNIST [22] and CIFAR-10/100 [23] used in [24; 25; 26; 27; 28]. Moreover, ASR simultaneously solves classification and segmentation tasks that add additional complexity on top of standard classification/regression tasks, de facto standard in many prior FL and DP works[2]. In natural language processing (NLP) data heterogeneity is defined by diverse user language characteristics while ASR broadens NLP data by including, e.g., acoustic, recording, and environment properties.

While large over-parameterized neural networks are becoming ubiquitous in many domains, e.g., computer vision [29; 30], NLP [31; 32] and ASR [2; 33; 5; 4; 3; 34], the same proliferation of large models has not been observed in FL and DP research [35; 36; 37] due to (i) communication complexity of FL [38] and (ii) difficulty of training large-scale models with DP [39; 40; 41][3]. However, practical model sizes are increasing over time, including ASR models [44]. Moreover, prior works reported several stark differences in training large-scale models compared to smaller models in the centralized training [45; 46; 47]. Some works [48] support a hypothesis that optimization of larger models is simpler: e.g., distillation from a large model into a small model is still the dominant method in the research community to train small models [49]. To fill the gap in understanding large-scale models in the context of FL and DP, and to alleviate the optimization issues related to training smaller models, we solely study large-scale models in this paper.

In this work, we aim to push for a collaborative step towards research in large-scale FL for production-competitive models, widespread ASR as a benchmark in FL and DP, and study the difficulties that arise in this setup. We observe several challenges when applying results from prior works in conventional FL research [35; 50] towards large-scale production-ready ASR models — architectures that use transformer [51] and conformer [2] blocks and are optimized using Connectionist Temporal Classification (CTC) [52] or CTC with attention encoder-decoder (AED) [53] loss. Through analyzing these further, we draw insights into the fundamental differences that arise out of model sizes and switching to ASR task:

- **Adaptive optimizers** Some adaptive optimizers perform better than others because they induce smoothness over the gradient subspace and/or reduce heterogeneity among FL client updates. Adaptive optimizers for FL with large-scale transformer models are necessary to match centrally trained models.
- **Robustness** While FL is sensitive to hyper-parameters of some optimizers, we are able to find a robust optimizer setting that is applicable to other data, e.g. the training recipe transfers out of the box from English data to French and German data.
- **Seed models** FL models can achieve nearly optimal performance for training both from scratch and from a seed model centrally pre-trained even on out-of-domain data. The discrepancy between different models reduces as the cohort size increases.
- **Data heterogeneity** The issue of data heterogeneity is alleviated with increased cohort size and adaptive optimizers. Also, transformer models appear to be less susceptible to data heterogeneity than conformer models.
- **FL with DP** While training transformer-based ASR models using FL with DP is fundamentally difficult due to the adverse effect of DP noise, this can be mitigated by revisiting per-layer clipping, especially when used together with layer-wise adaptive optimizers such as LAMB [45] and LARS [54].

[2]We wish to underscore that our critique proposes to set priority for the development of benchmarks on image segmentation tasks in FL and DP communities to approach realistic cases.

[3]Recent works [42; 43] showed that it is possible to fine-tune large language models centrally with hundreds of millions of parameters with DP. The DP noise does not affect training efficiency if gradients are low rank.

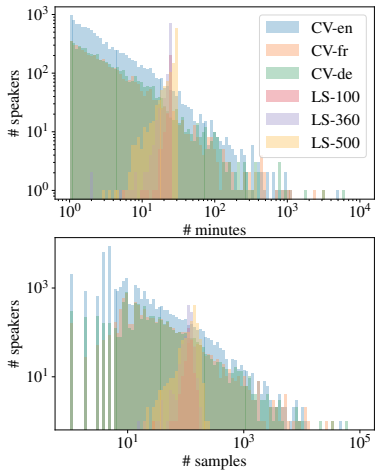

Figure 1: Speakers distribution in LibriSpeech (LS) and Common Voice (CV) train data: number of minutes per speaker (top) and number of samples per speaker (bottom).

Table 1: Speaker statistics for LibriSpeech (LS) and Common Voice (CV) train sets and their subsets.

Subset	# hours	# speakers	# minutes per speaker			
			mean	std	min	max
LS-100	100.6	251	24.1	2.7	5.5	25.2
LS-360	363.6	921	23.7	3.2	1.9	25.3
LS-500	496.9	1,166	25.6	5.9	3.0	30.3
LS-860	860.5	2,087	24.7	5.1	1.9	30.3
LS-960	961.1	2,338	24.7	4.9	1.9	30.3
CV-en-train	1593.7	34,753	2.8	32.7	0.02	5,049.6
CV-en-train-10	149.5	3,475	2.6	17.3	0.03	755.1
CV-en-train-90	1444.2	31,278	2.8	34.0	0.02	5,049.6
CV-en-train-05	79.5	1,737	2.7	15.8	0.03	508.3
CV-en-train-95	1514.2	33,016	2.7	33.4	0.02	5,049.6
CV-fr-train	727.9	6,856	6.4	57.2	0.04	3081.2
CV-fr-train-10	47.6	685	4.2	13.6	0.07	235.1
CV-fr-train-90	680.3	6,171	6.6	60.2	0.04	3081.2
CV-de-train	852.8	7,127	7.2	89.2	0.03	6249.9
CV-de-train-10	52.2	712	4.4	11.4	0.04	120.8
CV-de-train-90	800.6	6,415	7.5	94.0	0.03	6249.9

This paper summarizes results from our prior works [55; 56]. We refer the readers to these for a more in-depth analysis and additional discussion on the empirical results presented in this work.

2 Empirical Setup

Datasets. We perform experiments using two datasets of audio-transcription pairs: LibriSpeech (LS) [57] and Common Voice v13.0 (CV) [58]. These two datasets are read speech but differ in other properties, like the source of text (books for LS vs Wikipedia for CV), data diversity, noise conditions, speaker variation, and speaker distribution (see Figure 1 and Table 1). Since different languages vary in terms of ASR difficulty, for CV, we consider 3 languages: English, French, and German. One scenario that is of special interest for practical FL is to train a seed model centrally on a small dataset and fine-tune the seed model using FL on a larger dataset from another domain (e.g. seed model from a small subset of LS is fine-tuned with FL on CV). See Section 4 in [56] and Section 4 in [55] for more details on the datasets used.

Models. We discuss results with respect to two models used extensively in ASR:

- M1 from [56]: A vanilla encoder-based transformer model trained with the CTC loss [52]; we start our experimentation with the state-of-the-art model on *LS-100* from [59] with pre-LayerNorm and $\sim$ 255M parameters. Refer to Section 4 in [56] for more details on M1 setup and experiments.
- M2 from [55]: A conformer encoder & bidirectional transformer decoder model with $\sim$ 120M parameters trained using the CTC-AED loss. This model is derived from the WeNet [60] model config[4]. Refer to Section 4 in [55] for more details on M2 setup and experiments.

Apart from the difference in the model size, the other characteristics to be noted include:

- *layer characteristics:* while M1 only comprises vanilla attention layers, M2 contains both convolution and attention layers;
- *layer normalization:* both models utilize pre-layer normalization since we find that FL training fails to converge consistently when using post-layer normalization;
- *training losses:* while M1 is trained using CTC loss alone, M2 uses CTC-AED loss; and
- *tokenizer and stride:* while M1 is using 30ms stride and character tokenizer, M2 uses 60ms stride and SentencePiece [61] tokenizer with byte-pair encoding (BPE) subword units.

[4]Link to the configuration file: https://github.com/wenet-e2e/wenet/blob/main/examples/librispeech/s0/conf/train_conformer_bidecoder_large.yaml

FL Training FL introduces several challenges into the model training, e.g. heterogeneous data [62; 26; 25], scaling laws for large cohort training [27], and convergence rate due to local training [63; 64; 65]. In this paper, we focus on cross-device FL [24] where only a small fraction of users (clients) participate in each central iteration and thus users cannot maintain a state across rounds, e.g., local optimizer state or SCAFFOLD [66] state. The number of users participating in training in each central iteration is termed *cohort size* or L. We simulate FL by considering every speaker in the data as a separate user. Unless noted otherwise, we restrict the number of central steps to $T = 2$k. Moreover, for *practical* FL with DP we are limited by the total privacy budget that we can spend on hyper-parameter tuning, because it incurs additional overhead in terms of the privacy, and communication and computation cost [67; 68]. We use various combinations of local and global optimizers and learning rate schedules. Performance of models is measured by *word error rate (WER)* on the test data. For more details, see [55; 56].

3 Key Empirical Findings

In this section, we expand on the several characteristics of FL for ASR (FL4ASR) that stand out in terms of differences between: (i) FL4ASR models differing in sizes, architecture, and losses, i.e., the differences among M1 and M2, and (ii) prior works in conventional FL research vs. our observations for FL4ASR for M1 and/or M2. Also, we establish the *first benchmark* for FL with DP in ASR.

3.1 Adaptive Optimizers

Even within central training frameworks, it is well known that transformer models training poses unique challenges related to optimization, which can be tackled using adaptive optimizers as well as heuristics such as gradient clipping and learning rate scheduler among others [69; 70; 8]. Concurrently, similar observations have been made in the context of ASR training using transformer and conformer models [69; 60]. Similarly, the use of adaptive optimizers is also necessary in FL as it is claimed to help alleviate the negative impact of heterogeneous data on optimization efficiency and convergence speeds [35]. ASR provides a rich set of benchmarks with realistic, heterogeneous data, and competitive performance can be achieved by different models that vary in size, basic components (transformers or conformers), and formulation of the loss function (e.g. CTC or CTC with AED).

While prior works consider adaptive optimizers, our analysis gives us a deeper insight into the implications of using adaptive optimizers specifically towards ASR. In particular, when training FL from scratch, the usability of second-order estimates of central optimizers between different aggregation rounds (i.e., different samples of heterogeneous clients) can be interpreted in terms of domain shift, i.e., the second-order estimates gathered from one batch of clients in an aggregation round prove to be uninformative for making adaptive adjustments to the learning rate for a subsequent batch of heterogeneous clients. In our experiments, we find this domain shift between the gathered second-order estimates and subsequent client updates is especially dominant at the start of the FL training. However, we demonstrate that it can be mitigated by using one or more of the following techniques: (i) using layer-wise adaptive optimizers, (ii) increasing the cohort size during an aggregation round, (iii) initializing FL with a seed model, or (iv) starting training with SGD as central optimizer and switching to adaptive optimizers after a certain number of central iterations. We posit that adaptive optimizers have an impact by reducing the heterogeneity among client updates by inducing smoothness over gradient subspace. It can be observed empirically in terms of (i) an increase in cosine similarity among updates from different aggregation rounds, Figure 2 (left), (ii) an increase in cosine similarity among clients within an aggregation round, Figure 2 (right).

Our results also indicate that FL optimization appears more difficult with conformer models than with transformer models (e.g. compare Table 6 in [55] and Table 10 in [56]), although the transformer models are larger (in our case, the transformer model is approximately twice as large as the conformer model). Our analysis confirmed that this does not seem to be due to the AED component of the loss function used in most experiments with conformer models (see Table 3 in [55]). Also, [73] argued that transformers are less susceptible to distribution shifts and are thus especially well suited to deal with heterogeneous data in FL. Therefore, it may likely be the case that the transformer architecture is less susceptible to heterogeneous data also in the domain of ASR. See Section 3.4 for more discussion on the impact of data heterogeneity on FL for ASR.

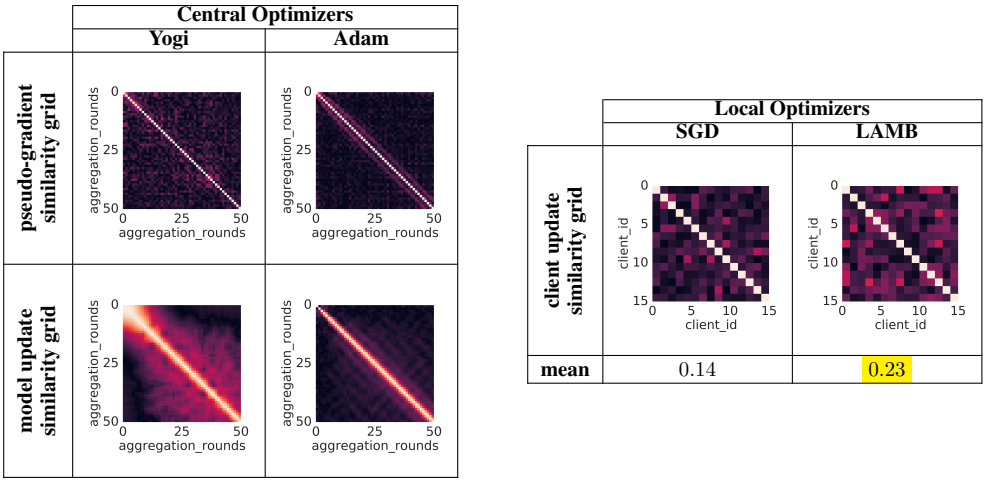

Figure 2: **Left:** Overlap (lighter color means higher cosine similarity) among *central* model updates for Yogi [71] and Adam [72] generated for first 50 aggregation rounds for the 7^{th} encoder layer (conv) with $\sim$ 3M parameters in the M2 model. **Right:** Overlap (lighter color means higher cosine similarity) among *client* model updates for SGD and LAMB [45] (central optimizer is Yogi) generated for 15 clients in an aggregation round for the 7^{th} encoder layer (conv) with $\sim$ 3 million parameters in the M2 model.

For transformer models, we found a number of central optimizers work comparably well (see Tables 8 and 9 in [56]); this provides further support to the observation that transformer models in FL for ASR are easier to train and more robust than conformer models of smaller size.

3.2 Seed Models

In FL, there are typically some data available on the server, which are often from a different distribution than the actual user data stored locally on user devices, and the amount of server data is often orders of magnitude smaller than that of user data. To improve the quality of trained FL models, the server data can be used to train a seed model, which can subsequently be used as a warm start initialization for fine-tuning with FL using the data stored locally on user devices [74; 75; 76; 77; 18]. The server data can also be used to tune hyper-parameters because tuning hyper-parameters with FL incurs significant privacy, communication, and computation costs.

We want to emphasize that ASR provides a suitable framework for studying the impact of seed models because there are several datasets from different sources and with different properties (e.g. in terms of the amount of data available for each user, acoustic variability of audio data, or the distribution of the content). For example, one practical scenario is to use a limited amount of data from one dataset (e.g. LS with $\sim$ 100h of audio) to train a seed model, and subsequently train the FL model on another, larger dataset (e.g. CV with $\sim 1,600$h of audio). Also, there is a shift in distribution between the central and FL data. This is a scenario we choose to use for experiments with DP (see Section 3.5).

Given that the use of seed models improves the final outcome of FL practically in every case [74; 75; 76; 77; 18; 78; 19], it is somewhat surprising that most prior works on FL do not study the impact of seed models on FL performance. Several of these prior works focused specifically on the use of seed models in FL for ASR [78; 19; 18].

Our results confirm the usefulness of starting with seed models (see Figures 3 and 4, and Tables 3, 4, 10 in [56]; Tables 3, 5, 6 in [55]): seed models improve performance of FL on ASR regardless of the model type, cohort size, language, or the combination of optimizers used for local and central training. Furthermore, we have shown that seed models improve FL performance even when the seed model comes from another domain (e.g. even a seed model trained on a small subset of LS improves the results of FL training on CV) and in some cases out-of-domain data could be preferable for the seed model training (e.g. LS seed for CV data). The latter we attribute to better generalization to longer audio/text sequences as LS has longer audio duration compared to CV. For example, for LS

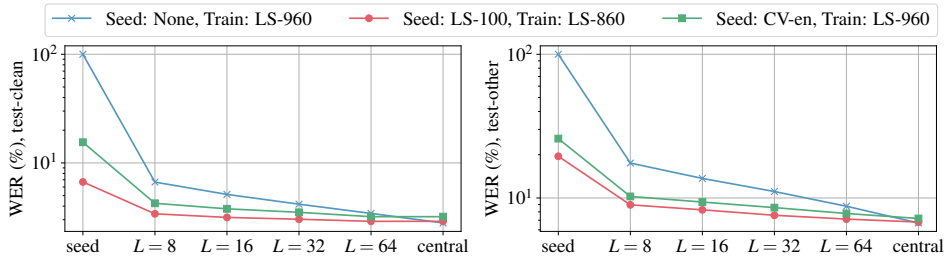

Figure 3: Comparison of WERs between central training and FL, and the impact of the cohort size L and seed models for M1 models trained on LS. We use exponential decay for central LR starting at $t = 1,000$, decay rate 0.6, and transition steps 500 (w/o seed model) or 250 (w/ seed model) with $T = 2$k total central steps and 10 local epochs. Local (central) LR is 0.4 (0.006) (w/o seed model) or 0.2 (0.003) (w/ seed model). Detailed results can be found in [56], Table 3 of Appendix.

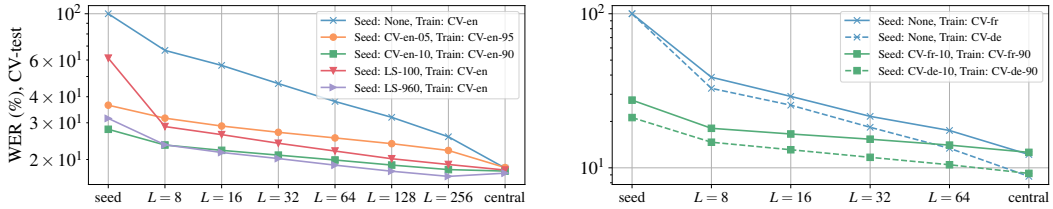

Figure 4: Comparison of word error rates (WERs) between central training and FL, and the impact of the cohort size L and seed models for M1 models trained on CV: English (left) and French/German (right). We use exponential decay for central LR starting at $t = 1{,}000$ (w/o seed model) or 750 (w/ seed model), decay rate 0.6, and transition steps 500 (w/o seed model) or 750 (w/ seed model) with $T = 2$k total central steps and 10 local epochs. Local (central) LR is 0.4 (0.006) (w/o seed model) or 0.2 (0.002) (w/ seed model). Detailed results and parameter settings can be found in [56], Tables 4 and 10 of Appendix.

data the best seed is in-domain despite CV has more data as CV seed suffers from generalization to longer sequences [7].

Recent work [18] focused on a small Seq2Seq model with CNN encoder and RNN decoder trained with joint CTC-attention objective using a sub-word tokenizer. The authors argued that it is "nearly impossible to train an E2E ASR model from scratch in a realistic FL setup" and proposed to use an additional training step over a small batch of held-out data on the server, after the FL model update. In contrast, our results show that FL training from scratch for *large* models is viable with a proper optimizer and without the need for additional training steps on server data (see Figures 3 and 4, Tables 3, 4 in [56] and Table 6 in [55]).

3.3 Cohort Size and Number of Local Epochs

Due to the communication complexity of FL, it is necessary that the training uses a moderate number of central iterations [74; 79]. This necessitates that we maximize the utility of each central iteration. One of the most straightforward ways of doing this is to increase *cohort size* and the number of local epochs. Good scalability of FL to large cohorts is critical also for ensuring strong privacy guarantees for FL using DP (see Section 3.5) and that is why some of the real-world FL deployments use cohort sizes as large as 150k [80]. However, some prior FL studies implied that scaling FL to large cohorts may be challenging [27].

Most prior works on FL for ASR did not provide an in-depth study of the impact of the cohort size on the quality of obtained models and instead used one or two cohort sizes for all the experiments. In contrast, we have analyzed a range of cohort sizes in a variety of scenarios with a varying initialization (with or without a seed model), dataset, model, loss function, and language (see Figures 3 and 4, Tables 3, 4 in [56] and Table 6 in [55]). In all cases, the larger the cohort size the better the performance within the fixed number of central iterations of 2k. This is expected since the amount

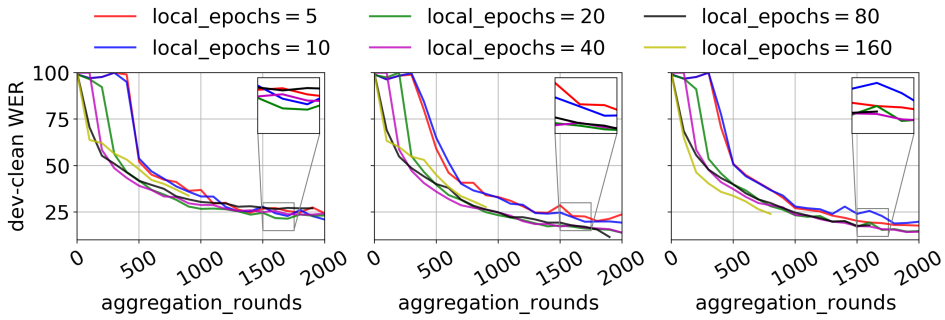

Figure 5: The effect of the number of local epochs on the convergence rate and performance for M2 model trained on (left) *LS-100*, (middle) *LS-100* and *LS-360*, and (right) *LS-960* data. A higher number of local epochs corresponds to faster convergence but gives diminishing returns (see green vs black lines) with a possible plateau on higher WER. This is consistent with observation in [81], i.e., diminishing benefits from increasing local epochs.

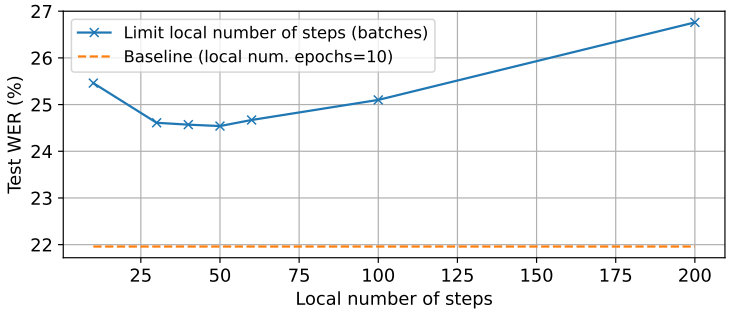

Figure 6: Comparison of WER for FL training of M1 model between local number of steps (solid) and local number of epochs (dashed). Training is done on *CV-en-train* with a seed model pre-trained centrally on *LS-100*. The cohort size is $L = 64$, total number of central steps is $T = 2$k, and all other parameters are set the same as in the corresponding configuration in Figure 4.

of data seen by a model in training grows linearly with the cohort size for a fixed number of central iterations.

Our results in Figure 5 confirm that increasing the number of local epochs generally speeds up the convergence of FL, although eventually, the large number of local epochs precludes finding competitive models and the benefits diminish as also discussed in [81]. We also tried to limit the number of steps on each client (as opposed to the number of epochs), but we saw degradation in the quality of the models in contrast to the observations of [26] as seen in Figure 6.

3.4 Heterogeneous Data

Numerous researchers argued that the heterogeneous data pose a challenge for FL [62; 26; 38] and also specifically for FL in ASR [82; 20; 15; 18]. In most prior work on data heterogeneity in FL, the focus is either on synthetic datasets or real-world datasets such as MNIST [22] and CIFAR [23] where the data are distributed among the devices using a prior distribution that ensures that the data are heterogeneous. Using synthetic data and real data with a synthetic distribution across user devices can provide useful insights into the challenges for FL in the presence of data heterogeneity. However, the good news is that ASR datasets such as LS and especially CV are inherently heterogeneous and provide user IDs for each audio-transcription pair. This provides a realistic scenario for testing the impact of data heterogeneity on FL and the best ways to alleviate its effects. However, we admit that ASR data may not exhibit all types of heterogeneity that can be observed in the real world.

Our results for transformer models (M1) show that distributing data uniformly and randomly across users indeed improves performance as seen for all datasets, cohort sizes, and seed models in Figure 7. However, the impact is marginal and it decreases with increasing cohort size. On the contrary, the

training with conformer models (M2) leads to the observation that training using Adam optimizer only converges for iid data, especially with a smaller ($L = 5$) cohort size; it converges for larger cohort sizes as discussed in Section 3.1. One of the plausible explanations for this discrepancy is that the transformers were claimed to be relatively less susceptible to heterogeneous data [73].

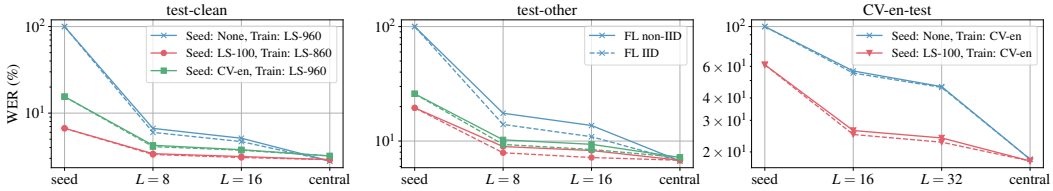

Figure 7: Impact of randomizing the distribution of data across users for LS (left, middle) and CV (right) measured by WER for M1 model. Parameter settings are described in Figure 3 for LS and Figure 4 for CV. While the original training data are non-IID (solid), IID (dashed) versions of *LS-960*, *LS-860* and *CV-en-train* are created by choosing a user id uniformly and randomly from the set of user ids for each data point in the corresponding dataset. Details are in [56], Tables 5 and 6.

3.5 Differential Privacy

While FL on its own provides only limited privacy guarantees [83; 84; 85; 86; 87], it can be combined with differential privacy (DP) [12] and secure aggregation [13; 14] to provide strong privacy guarantees for users (or clients) while training high-quality models [50]. FL on its own provides limited privacy even in the context of ASR [88; 19]. However, to the best of our knowledge, no prior work exists that incorporates DP into FL for ASR. We argue that using DP in FL for ASR is hard but feasible, and provide the *first benchmarks for FL with DP for ASR*.

Algorithm 1: Federated learning with differential privacy (marked as red) for transformer model

Inputs: Initial model state $\boldsymbol{\theta}^0$ (either randomly initialized or pre-trained on server data), central steps T, central optimizer opt, clients sampling rate q, local steps T_l, local optimizer opt_l, clipping function $\text{clip}(\boldsymbol{v}, C) = \boldsymbol{v} \cdot \min\left(1, \frac{C}{||\boldsymbol{v}||}\right)$, local clipping C_l, DP clipping C and DP noise σ.

Result: ASR model $\boldsymbol{\theta}^T$

Initialize central optimizer opt
for $t = \overline{1, T}$ **do**
  Sample every client with probability q to form a subset $\mathcal{N}^t$ of clients from all clients $\mathcal{N}$ ($|\mathcal{N}| = N$)
  `// For practical implementation we fix the size of the cohort` $\mathcal{N}^t$ `to` L `throughout the training.`
  for $n = \overline{1, |\mathcal{N}^t|}$ `in parallel` **do**
    Initialize local model $\boldsymbol{\theta}_n^{(t,0)} \leftarrow \boldsymbol{\theta}^{t-1}$ and local optimizer opt_l
    for $t_l = \overline{1, T_l}$ **do**
      `// We also use local epochs instead of steps: then this loop has different number of steps per client.`
      Sample train mini-batch $\boldsymbol{B}_n^{t_l} \in \mathcal{D}_{\mathcal{N}_n^t}$ and compute gradient estimate $\boldsymbol{g}_n^{(t,t_l)}(\boldsymbol{B}_n^{t_l}; \boldsymbol{\theta}_n^{(t,t_l-1)})$
      Clip gradients $\boldsymbol{g}_n^{(t,t_l)} \leftarrow \text{clip}(\boldsymbol{g}_n^{(t,t_l)}, C_l)$ and update a local model $\boldsymbol{\theta}_n^{(t,t_l)} \leftarrow \text{opt}_l(\boldsymbol{g}_n^{(t,t_l)})$
    Compute client's delta $\boldsymbol{\Delta}_n^t = \boldsymbol{\theta}_n^{(t,T_l)} - \boldsymbol{\theta}_n^{(t,0)} = \boldsymbol{\theta}_n^{(t,T_l)} - \boldsymbol{\theta}^{t-1}$
    Clip client's delta $\boldsymbol{\Delta}_n^t \leftarrow \text{clip}(\boldsymbol{\Delta}_n^t, C)$
    Add Gaussian noise to client's delta $\boldsymbol{\Delta}_n^t \leftarrow \boldsymbol{\Delta}_n^t + \mathcal{N}(0, I\sigma^2 qN)$
  Compute central model's pseudo-gradient $\boldsymbol{g}^t = \boldsymbol{\Delta}^t = \frac{1}{|\mathcal{N}^t|} \sum_{n=1}^{|\mathcal{N}^t|} \boldsymbol{\Delta}_n^t$
  Update the central model $\boldsymbol{\theta}^t \leftarrow \text{opt}(\boldsymbol{g}^t)$

To incorporate user-level DP into FL, there are two additional steps compared to vanilla FL (see Algorithm 1): (i) clipping user deltas $\boldsymbol{\Delta}_n^t$ so that they have bounded L_2 norm $\|\boldsymbol{\Delta}_n^t\|_2 \leq C$ at every central training step t; (ii) addition of Gaussian noise $\mathcal{N}(0, I\sigma^2 qN)$ to the clipped deltas before sending them to the server where $q = L/N$, L is cohort size, and N is the total number of users (population size). DP guarantees can be quantified using ϵ and δ [12]; δ is typically a small number between 10^{-9} to 10^{-6} and $\epsilon < 10$ to be practical, ideally between 1 and 2.

Table 2: Results for FL with DP and a M1 model pre-trained on *LS-100* (∼100h) used as central data and afterwards fine-tuned on *CV-en-train* (∼1.6k hours) used as clients data. We report added noise $\mathcal{N}(0, I\sigma^2 Nq)$ per client and CV dev and test WERs (%) for two clipping variants with clipping bound C: global and per-layer "uniform" ("dim"). The total number of users is N, the expected number of users sampled per central step is $L = qN$, and the number of central steps is T. We set $\delta = 10^{-9}$ and report ϵ for which (ϵ, δ)-DP holds for given L and N using the moments accountant of [50]. For scaling L and N where it is practically intractable to run model training (marked "-"), we extrapolate (ϵ, δ)-DP assuming the training dynamic remains unchanged and thus similar WER could be obtained. Central training gives 14.7%/17.8% WER on dev/test. Extended results see in [56]. ϵ should be below 10 to be practically useful (marked with blue).

z	σ $\cdot 10^{-8}$	C	L	N	$q = L/N$	T	ϵ	Renyi order	global clipping		per-layer clipping: uniform (dim)	
									dev WER	test WER	dev WER	test WER
-	-	-	0	34,753	0	0	0	-	54.7	61.2	54.7	61.2
0.03072	30.0	0.01	1,024	34,753	0.0295	2,006	$1.1{\cdot}10^{6}$	1.1	-	-	25.2 (24.2)	29.3 (28.2)
0.3072	30.0	0.01	10,240	347,530	0.0295	2,006	$3.7{\cdot}10^{2}$	1.1	-	-	-	-
1.536	30.0	0.01	51,200	1,737,650	0.0295	2,006	$6.5{\cdot}10^{0}$	7.0	-	-	-	-
0.02048	20.0	0.01	1,024	34,753	0.0295	2,006	$2.6{\cdot}10^{6}$	1.1	-	-	23.7 (22.6)	27.6 (26.5)
1.024	20.0	0.01	51,200	1,737,650	0.0295	2,006	$1.3{\cdot}10^{0}$	4.0	-	-	-	-
2.048	20.0	0.01	102,400	3,475,300	0.0295	2,006	$4.5{\cdot}10^{0}$	9.0	-	-	-	-
0.01024	10.0	0.01	1,024	34,753	0.0295	2,006	$1.1{\cdot}10^{7}$	1.1	-	-	**21.3 (20.1)**	**25.0 (23.7)**
0.512	10.0	0.01	51,200	1,737,650	0.0295	2,006	$7.2{\cdot}10^{1}$	1.5	-	-	-	-
1.024	10.0	0.01	102,400	3,475,300	0.0295	2,006	$1.3{\cdot}10^{1}$	4.0	-	-	-	-
2.048	10.0	0.01	204,800	6,950,600	0.0295	2,006	$4.5{\cdot}10^{0}$	9.0	-	-	-	-
0.003072	3.0	0.01	1,024	34,753	0.0295	2,006	$1.2{\cdot}10^{8}$	1.1	27.0	31.1	**17.9 (17.1)**	**21.2 (20.4)**
0.3072	3.0	0.01	102,400	3,475,300	0.0295	2,006	$3.7{\cdot}10^{2}$	1.1	-	-	-	-
0.6144	3.0	0.01	204,800	6,950,600	0.0295	2,006	$4.2{\cdot}10^{1}$	2.0	-	-	-	-
0.6144	3.0	0.01	204,800	69,506,000	0.00295	2,034	$7.2{\cdot}10^{0}$	3.0	-	-	-	-
0.6144	3.0	0.01	204,800	695,060,000	0.000295	3,390	$3.7{\cdot}10^{0}$	6.0	-	-	-	-
0.001024	1.0	0.01	1,024	34,753	0.0295	2,006	$1.1{\cdot}10^{9}$	1.1	22.9	26.7	16.2 (16.0)	19.5 (19.3)
0.2048	1.0	0.01	204,800	6,950,600	0.0295	2,006	$1.1{\cdot}10^{3}$	1.1	-	-	-	-
0.2048	1.0	0.01	204,800	69,506,000	0.00295	2,034	$2.7{\cdot}10^{2}$	1.1	-	-	-	-
0.2048	1.0	0.01	204,800	695,060,000	0.000295	3,390	$9.4{\cdot}10^{1}$	1.3	-	-	-	-
-	0	0.01	1,024	34,753	0.0295	2,000	inf	-	15.7	18.9	15.9	19.1
-	0	1.0	1,024	34,753	0.0295	2,000	inf	-	15.7	18.9	15.9	19.1

Figure 8 shows the word error rate (WER) measured on *CV-en-test* set for different values of DP noise and different clipping algorithms; see Section 3 and Section 5.3 in [56] for details. Although prior work [37] suggested that the per-layer clipping does not improve performance of FL with DP, our results show that per-layer clipping performs considerably better than global clipping: with the same DP noise $\sigma = 3 \cdot 10^{-8}$ we are able to closely match the model trained without DP noise ($\sigma = 0$) with only a small WER degradation (from 19.1% to 21.2% WER) while guaranteeing (7.2, 10^{-9})-DP assuming the training effectiveness remains the same if we extrapolate scaling to ∼70M clients with the cohort size of ∼200k. Moreover, we can now increase DP noise up to $\sigma = 10^{-7}$ getting 25.0% WER with (4.5, 10^{-9})-DP by scaling only to ∼7M clients with the cohort size of ∼200k (see Table 2). The latter is a realistic scenario even for mid/low resource languages. We can further reduce WER by ∼1% for the same (ϵ, δ)-DP guarantee if we apply per-layer clipping based on the layer dimension ("dim"), Table 2.

Interestingly, for both French and German we observe that per-layer clipping is not as effective as for English and we get only marginal improvements over global clipping. We have checked that the seed model quality and the seed model being out-of-domain are not the sources of this discrepancy in results between languages. Further investigation indicated that there is a discrepancy in gradients balance across layers for the central model training for English, French, and German, which is likely due to the differences between the languages. For more details, we refer a reader to [56].

4 Conclusion

ASR provides a valuable benchmark for (private) federated learning (FL). Datasets in this domain are large, separated by users, and represent heterogeneity of the kind often seen in real private FL settings. With the possible exception of language modeling, benchmarks typically used in the private FL research community do not satisfy these properties, making them less suitable for deriving conclusions that would translate to practical deployments. For example, the task of adapting a model trained (centrally) on LibriSpeech data to work on Common Voice data in the federated setting is

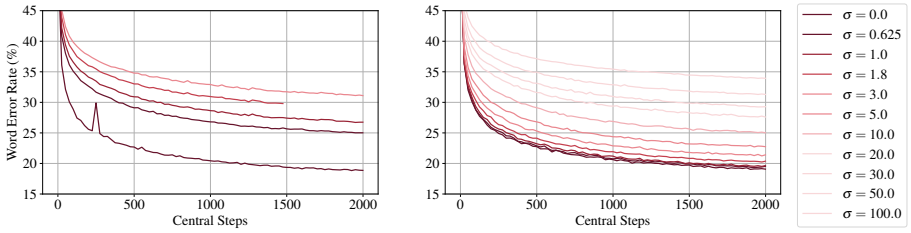

Figure 8: Word error rate (WER) of M1 model measured on *CV-en-test* set for different values of DP noise σ (scale is set to 10^{-8}). We apply clipping of 10^{-2} either globally (left, Algorithm 1) or per-layer (right, "uniform" in [56], Section 3) with $T = 2$k central steps, 10 local steps, and $L = 1{,}024$ cohort size. The seed model is trained on *LS-100*.

an appropriate candidate for benchmarking FL and FL with differential privacy (DP). In this paper, we argue that ASR benchmarks provide insights into all critical aspects of the practical deployment of private FL in challenging problem domains, including the impact of adaptive optimization, data heterogeneity, cohort size, and other FL parameters, seed models, and DP. With a *practical* number of central aggregations, we have been able to train large transformer and conformer FL models that are nearly optimal even with heterogeneous data, a seed model from another domain, or no pre-trained seed model. We show that even with large transformer models, FL with DP can achieve user-level $(7.2, 10^{-9})$-DP (resp. $(4.5, 10^{-9})$-DP) with a 1.3% (resp. 4.6%) absolute drop in the word error rate for extrapolation to high (resp. low) population scale.

Acknowledgement

We would like to thank (in alphabetical order within the group):

- Samy Bengio, David Grangier, Filip Granqvist, Navdeep Jaitly, and Vojta Jina for essential general discussion on the paper throughout all stages;
- Pierre Ablin and Dan Busbridge for discussion on scaling laws;
- Audra McMillan and Congzheng Song for discussion on differential privacy;
- Shuangfei Zhai for discussion on transformer stability and behavior of gradient norms;
- Dan Busbridge, Ronan Collobert, Roger Hsiao, Navdeep Jaitly, Audra McMillan, and Barry Theobald for the helpful feedback on the initial drafts of the work;
- Hassan Babaie, Cindy Liu, Rajat Phull, and the wider Apple infrastructure team for assistance with developing scalable, fault-tolerant code.

References

[1] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end ASR: From supervised to semi-supervised learning with modern architectures," in *ICML 2020 Workshop on Self-supervision in Audio and Speech*, 2020. [Online]. Available: https://openreview.net/forum?id=OSVxDDc360z 1

[2] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2020. 1, 2

[3] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi, "SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network," *arXiv preprint arXiv:2104.02133*, 2021. 1, 2

[4] T. Likhomanenko, Q. Xu, V. Pratap, P. Tomasello, J. Kahn, G. Avidov, R. Collobert, and G. Synnaeve, "Rethinking Evaluation in ASR: Are Our Models Robust Enough?" *arXiv preprint arXiv:2010.11745*, 2020. 1, 2

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 12 449–12 460, 2020. 1, 2

[6] L. Lugosch, T. Likhomanenko, G. Synnaeve, and R. Collobert, "Pseudo-labeling for massively multilingual speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7687–7691. 1

[7] T. Likhomanenko *et al.*, "CAPE: Encoding relative positions with continuous augmented positional embeddings," *Neural Information Processing Systems (NeurIPS)*, 2021. 1, 6

[8] S. Zhai, T. Likhomanenko, E. Littwin, D. Busbridge, J. Ramapuram, Y. Zhang, J. Gu, and J. M. Susskind, "Stabilizing transformer training by preventing attention entropy collapse," in *International Conference on Machine Learning (ICML)*, 2023. 1, 4

[9] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman *et al.*, "Seamlessm4t-massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023. 1

[10] D. Busbridge, J. Ramapuram, P. Ablin, T. Likhomanenko, E. G. Dhekane, X. Suau, and R. Webb, "How to scale your ema," *arXiv preprint arXiv:2307.13813*, 2023. 1

[11] J. Konečný, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," *CoRR*, vol. abs/1511.03575, 2015. [Online]. Available: http://arxiv.org/abs/1511.03575 1

[12] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014. 1, 8

[13] K. A. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for federated learning on user-held data," *CoRR*, vol. abs/1611.04482, 2016. [Online]. Available: http://arxiv.org/abs/1611.04482 1, 8

[14] K. Talwar, S. Wang, A. McMillan, V. Jina, V. Feldman, B. Basile, Á. Cahill, Y. S. Chan, M. Chatzidakis, J. Chen, O. Chick, M. Chitnis, S. Ganta, Y. Goren, F. Granqvist, K. Guo, F. Jacobs, O. Javidbakht, A. Liu, R. Low, D. Mascenik, S. Myers, D. Park, W. Park, G. Parsa, T. Pauly, C. Priebe, R. Rishi, G. Rothblum, M. Scaria, L. Song, C. Song, K. Tarbe, S. Vogt, L. Winstrom, and S. Zhou, "Samplable anonymous aggregation for private federated data analysis," *CoRR*, vol. abs/2307.15017, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2307.15017 1, 8

[15] D. Guliani, F. Beaufays, and G. Motta, "Training Speech Recognition Models with Federated Learning: A Quality/cost Framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. 1, 7

[16] W. Yu, J. Freiwald, S. Tewes, F. Huennemeyer, and D. Kolossa, "Federated learning in ASR: Not as easy as you think," in *Speech Communication; 14th ITG Conference*, 2021, pp. 1–5. 1

[17] D. Guliani, L. Zhou, C. Ryu, T.-J. Yang, H. Zhang, Y. Xiao, F. Beaufays, and G. Motta, "Enabling on-device Training of Speech Recognition Models with Federated Dropout," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 1

[18] Y. Gao, T. Parcollet, S. Zaiem, J. Fernandez-Marques, P. P. de Gusmao, D. J. Beutel, and N. D. Lane, "End-to-end Speech Recognition from Federated Acoustic Models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 1, 2, 5, 6, 7

[19] T. Nguyen, S. Mdhaffar, N. Tomashenko, J.-F. Bonastre, and Y. Estève, "Federated learning for ASR based on wav2vec 2.0," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. 1, 5, 8

[20] X. Cui, S. Lu, and B. Kingsbury, "Federated acoustic modeling for automatic speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6748–6752. 1, 7

[21] J. L. Kröger, O. H.-M. Lutz, and P. Raschke, *Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference*. Cham: Springer International Publishing, 2020, pp. 242–258. [Online]. Available: https://doi.org/10.1007/978-3-030-42504-3_16 2

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 2, 7

[23] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:18268744 2, 7

[24] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated Learning: Strategies for Improving Communication Efficiency," in *Neural Information Processing Systems (NeurIPS)*, 2016. 2, 4

[25] S. S. Azam, "Towards privacy and communication efficiency in distributed representation learning," Ph.D. dissertation, Purdue University, 2022. 2, 4

[26] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization," *Neural Information Processing Systems (NeurIPS)*, 2020. 2, 4, 7

[27] Z. Charles, Z. Garrett, Z. Huo, S. Shmulyian, and V. Smith, "On large-cohort training for federated learning," *Advances in neural information processing systems*, vol. 34, pp. 20 461–20 475, 2021. 2, 4, 6

[28] Z. Zhou, S. S. Azam, C. Brinton, and D. I. Inouye, "Efficient federated domain translation," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=uhLAcrAZ9cJ 2

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy 2

[30] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986. 2

[31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *Neural Information Processing Systems (NeurIPS)*, 2020. 2

[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019. 2

[33] D. Wu, B. Zhang, C. Yang, Z. Peng, W. Xia, X. Chen, and X. Lei, "U2++: Unified Two-pass Bidirectional End-to-end Model for Speech Recognition," *arXiv preprint arXiv:2106.05642*, 2021. 2

[34] S. Kim, A. Gholami, A. Shaw, N. Lee, K. Mangalam, J. Malik, M. W. Mahoney, and K. Keutzer, "Squeezeformer: An efficient transformer for automatic speech recognition," *Neural Information Processing Systems (NeurIPS)*, 2022. 2

[35] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive Federated Optimization," in *International Conference on Learning Representations (ICLR)*, 2021. 2, 4

[36] X. Zhang, X. Chen, M. Hong, Z. S. Wu, and J. Yi, "Understanding clipping for federated learning: Convergence and client-level differential privacy," in *International Conference on Machine Learning (ICML)*, 2022. 2

[37] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning Differentially Private Recurrent Language Models," in *International Conference on Learning Representations (ICLR)*, 2018. 2, 9

[38] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao, "Advances and open problems in federated learning," 2021. [Online]. Available: https://arxiv.org/abs/1912.04977 2, 7

[39] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, 2014, pp. 464–473. 2

[40] Y. Shen, Z. Wang, R. Sun, and X. Shen, "Towards understanding the impact of model size on differential private classification," *CoRR*, vol. abs/2111.13895, 2021. [Online]. Available: https://arxiv.org/abs/2111.13895 2

[41] F. Tramèr and D. Boneh, "Differentially private learning needs better features (or much more data)," *CoRR*, vol. abs/2011.11660, 2020. [Online]. Available: https://arxiv.org/abs/2011.11660 2

[42] X. Li, F. Tramer, P. Liang, and T. Hashimoto, "Large language models can be strong differentially private learners," in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=bVuP3ltATMz 2

[43] X. Li, D. Liu, T. B. Hashimoto, H. A. Inan, J. Kulkarni, Y.-T. Lee, and A. Guha Thakurta, "When does differentially private learning not suffer in high dimensions?" *Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 28 616–28 630, 2022. 2

[44] R. Botros, A. Gulati, T. N. Sainath, K. Choromanski, R. Pang, T. Strohman, W. Wang, and J. Yu, "Practical conformer: Optimizing size, speed and flops of conformer for on-device and cloud ASR," *arXiv preprint arXiv:2304.00171*, 2023. 2

[45] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh, "Large Batch Optimization for Deep Learning: Training BERT in 76 minutes," in *International Conference on Learning Representations (ICLR)*, 2020. 2, 5

[46] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research (JMLR)*, 2022. 2

[47] R. Eldan and Y. Li, "Tinystories: How small can language models be and still speak coherent english?" *arXiv preprint arXiv:2305.07759*, 2023. 2

[48] B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro, "Kernel and rich regimes in overparametrized models," in *Conference on Learning Theory*. PMLR, 2020, pp. 3635–3673. 2

[49] S. Stanton, P. Izmailov, P. Kirichenko, A. A. Alemi, and A. G. Wilson, "Does knowledge distillation really work?" *Neural Information Processing Systems (NeurIPS)*, 2021. 2

[50] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 308–318. 2, 8, 9

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," *Neural Information Processing Systems (NeurIPS)*, 2017. 2

[52] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *International Conference on Machine Learning (ICML)*, 2006. 2, 3

[53] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based Models for Speech Recognition," *Neural Information Processing Systems (NeurIPS)*, 2015. 2

[54] Y. You, I. Gitman, and B. Ginsburg, "Scaling SGD Batch Size to 32K for ImageNet Training," *CoRR*, vol. abs/1708.03888, 2017. [Online]. Available: http://arxiv.org/abs/1708.03888 2

[55] S. S. Azam, T. Likhomanenko, M. Pelikan, J. Silovsky *et al.*, "Importance of smoothness induced by optimizers in fl4asr: Towards understanding federated learning for end-to-end asr," *arXiv preprint arXiv:2309.13102*, 2023. 3, 4, 5, 6

[56] M. Pelikan, S. S. Azam, V. Feldman, J. Silovsky, K. Talwar, T. Likhomanenko *et al.*, "Federated learning with differential privacy for end-to-end speech recognition," *arXiv preprint arXiv:2310.00098*, 2023. 3, 4, 5, 6, 8, 9, 10

[57] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus based on Public Domain Audio Books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015. 3

[58] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222. 3

[59] T. Likhomanenko, Q. Xu, J. Kahn, G. Synnaeve, and R. Collobert, "slimIPL: Language-model-free iterative pseudo-labeling," *Conference of the International Speech Communication Association (INTERSPEECH)*, 2021. 3

[60] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "WeNet: Production oriented Streaming and Non-streaming End-to-End Speech Recognition Toolkit," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2021. 3, 4

[61] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Association for Computational Linguistics (ACL)*, 2016. 3

[62] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems*, I. Dhillon, D. Papailiopoulos, and V. Sze, Eds., vol. 2, 2020, pp. 429–450. 4, 7

[63] G. Malinovsky, K. Yi, and P. Richtárik, "Variance reduced ProxSkip: Algorithm, theory and application to federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15 176–15 189, 2022. 4

[64] S. Hosseinalipour, S. S. Azam, C. G. Brinton, N. Michelusi, V. Aggarwal, D. J. Love, and H. Dai, "Multi-stage hybrid federated learning over large-scale d2d-enabled fog networks," *IEEE/ACM transactions on networking*, vol. 30, no. 4, pp. 1569–1584, 2022. 4

[65] F. P.-C. Lin, S. Hosseinalipour, S. S. Azam, C. G. Brinton, and N. Michelusi, "Semi-decentralized federated learning with cooperative d2d local model aggregations," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3851–3869, 2021. 4

[66] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning (ICML)*, 2020. 4

[67] H. Wang, S. Sievert, S. Liu, Z. Charles, D. Papailiopoulos, and S. Wright, "ATOMO: Communication-efficient Learning via Atomic Sparsification," *Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018. 4

[68] S. S. Azam, S. Hosseinalipour, Q. Qiu, and C. Brinton, "Recycling Model Updates in Federated Learning: Are Gradient Subspaces Low-rank?" in *International Conference on Learning Representations (ICLR)*, 2022. 4

[69] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang *et al.*, "BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1519–1532, 2022. 4

[70] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin *et al.*, "Scaling vision transformers to 22 billion parameters," in *International Conference on Machine Learning (ICML)*, 2023. 4

[71] M. Zaheer, S. Reddi, D. Sachan, S. Kale, and S. Kumar, "Adaptive methods for nonconvex optimization," *Neural Information Processing Systems (NeurIPS)*, 2018. 5

[72] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014. 5

[73] L. Qu, Y. Zhou, P. P. Liang, Y. Xia, F. Wang, E. Adeli, L. Fei-Fei, and D. L. Rubin, "Rethinking architecture design for tackling data heterogeneity in federated learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 10 051–10 061. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.00982 4, 8

[74] Z. Xu, Y. Zhang, G. Andrew, C. Choquette, P. Kairouz, B. Mcmahan, J. Rosenstock, and Y. Zhang, "Federated learning of gboard language models with differential privacy," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 629–639. [Online]. Available: https://aclanthology.org/2023.acl-industry.60 5, 6

[75] B. Y. Lin, C. He, Z. Zeng, H. Wang, Y. Huang, M. Soltanolkotabi, X. Ren, and S. Avestimehr, "FedNLP: A research platform for federated learning in natural language processing," *CoRR*, vol. abs/2104.08815, 2021. [Online]. Available: https://arxiv.org/abs/2104.08815 5

[76] J. Stremmel and A. Singh, "Pretraining federated text models for next word prediction," *CoRR*, vol. abs/2005.04828, 2020. [Online]. Available: https://arxiv.org/abs/2005.04828 5

[77] H. Chen, C. Tu, Z. Li, H. Shen, and W. Chao, "On the importance and applicability of pre-training for federated learning," in *International Conference on Learning Representations (ICLR)*, 2023. [Online]. Available: https://openreview.net/pdf?id=fWWFv--P0xP 5

[78] J. Jia, J. Mahadeokar, W. Zheng, Y. Shangguan, O. Kalinli, and F. Seide, "Federated domain adaptation for ASR with full self-supervision," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2022. 5

[79] K. A. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, B. McMahan, T. V. Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," in *Proceedings of Machine Learning and Systems 2019, MLSys 2019, Stanford, CA, USA, March 31 - April 2, 2019*, A. Talwalkar, V. Smith, and M. Zaharia, Eds. mlsys.org, 2019. [Online]. Available: https://proceedings.mlsys.org/book/271.pdf 6

[80] M. Chatzidakis, J. Chen, O. Chick, E. Circlaeays, S. Gopalan, Y. Goren, K. Guo, M. Hesse, O. Javidbakht, V. Jina, K. Kalu, A. Katti, A. Liu, R. Low, A. McMillan, J. Meyer, S. Myers, A. Palmer, D. Park, G. Parsa, P. Pelzl, R. Rishi, M. Scaria, C. Sumanth, K. Talwar, K. Tarbe, S. Wang, and M. Yadav, "Learning iconic scenes with differential privacy," https://machinelearning.apple.com/research/scenes-differential-privacy, 2023, accessed: 2023-09-05. 6

[81] K. Mishchenko, G. Malinovsky, S. Stich, and P. Richtárik, "ProxSkip: Yes! Local Gradient Steps Provably lead to Communication Acceleration! Finally!" in *International Conference on Learning Representations (ICLR)*, 2022. 7

[82] D. Dimitriadis, K. Kumatani, R. Gmyr, Y. Gaur, and S. E. Eskimez, "A federated approach in training acoustic models," in *Interspeech*, 2020, pp. 981–985. 7

[83] F. Boenisch, A. Dziedzic, R. Schuster, A. S. Shamsabadi, I. Shumailov, and N. Papernot, "When the curious abandon honesty: Federated learning is not private," in *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2023, pp. 175–199. 8

[84] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=TatRHT_1cK 8

[85] S. Kariyappa, C. Guo, K. Maeng, W. Xiong, G. E. Suh, M. K. Qureshi, and H.-H. S. Lee, "Cocktail party attack: Breaking aggregation-based privacy in federated learning using independent component analysis," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 15 884–15 899. [Online]. Available: https://proceedings.mlr.press/v202/kariyappa23a.html 8

[86] M. Bertran, N. Martinez, A. Papadaki, Q. Qiu, M. Rodrigues, G. Reeves, and G. Sapiro, "Adversarially learned representations for information obfuscation and inference," in *International Conference on Machine Learning*. PMLR, 2019, pp. 614–623. 8

[87] S. S. Azam, T. Kim, S. Hosseinalipour, C. Joe-Wong, S. Bagchi, and C. Brinton, "Can we Generalize and Distribute Private Representation Learning?" in *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2022, pp. 11 320–11 340. 8

[88] N. Tomashenko, S. Mdhaffar, M. Tommasi, Y. Estève, and J.-F. Bonastre, "Privacy attacks for automatic speech recognition acoustic models in a federated learning framework," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6972–6976. 8