FEDERATED SELF-LEARNING WITH WEAK SUPERVISION FOR SPEECH RECOGNITION

Milind Rao *Gopinath Chennupati* *Gautam Tiwari* *Anit Kumar Sahu*
Anirudh Raju *Ariya Rastrow* *Jasha Droppo*

Amazon Alexa AI, U.S.A.

ABSTRACT

Automatic speech recognition (ASR) models with low-footprint are increasingly being deployed on edge devices for conversational agents, which enhances privacy. We study the problem of federated continual incremental learning for recurrent neural network-transducer (RNN-T) ASR models in the privacy-enhancing scheme of learning on-device, without access to ground truth human transcripts or machine transcriptions from a stronger ASR model. In particular, we study the performance of a self-learning based scheme, with a paired teacher model updated through an exponential moving average of ASR models. Further, we propose using possibly noisy weak-supervision signals such as feedback scores and natural language understanding semantics determined from user behavior across multiple turns in a session of interactions with the conversational agent. These signals are leveraged in a multi-task policy-gradient training approach to improve the performance of self-learning for ASR. Finally, we show how catastrophic forgetting can be mitigated by combining on-device learning with a memory-replay approach using selected historical datasets. These innovations allow for 10% relative improvement in WER on new use cases with minimal degradation on other test sets in the absence of strong-supervision signals such as ground-truth transcriptions.

Index Terms: Automatic Speech Recognition, Weak Supervision, Self Learning, Federated Learning

1. INTRODUCTION

On-device deployment of voice technologies enables use of conversational agents in settings without a reliable network connection to the cloud. It enables lower-latency responses by removing the need for utterances to be transmitted to the cloud for processing. Offline use, vehicular control, and healthcare are new use cases within this paradigm. When ASR is deployed on-device, models need to be adapted for specific acoustic or linguistic content specific to the deployment as well as temporal adaptation to distribution shifts in use across time. In this work, we look at continually and incrementally updating ASR models with resource constraints of memory and compute at the device in *federated* settings, i.e., privacy-enhancing features where (1) utterances are not transmitted to the cloud, (2) persistent storage of audio is not required, and (3) human ground-truth annotations of the audio need not be obtained.

Privacy-preserving machine learning [1] can enable learning from user data while mitigating privacy risks. Federated learning (FL) [2] is one of the most popular privacy-preserving learning frameworks which involves training models on-device, with data not leaving edge devices. In FL, multiple model updates from a number of participating devices are aggregated securely on a central server at every round. FL has been demonstrated to perform well in speech applications such as speech recognition [3], keyword spotting [4], and speaker verification [5] among others. Mixed centralized and federated training was done in [6] and layer-wise representation learning in [7]. However, the aforementioned works involve training a model from scratch instead of fine-tuning a well-trained model. In addition, previous works considered static data which does not change across rounds. Differently from previous work, we consider FL settings, where the model is initialized to a well-trained model and streaming data on devices, which are not persisted across rounds. In [8], authors look at domain adaptation of ASR in a federated setting. We additionally look at incorporating weak supervision to learn from alternate sources of feedback.

Semi-supervised learning (SSL) deals with training and improving ASR using unlabelled audio, such as the audio available at devices. Unsupervised approaches such as data2vec [9] or WavLM [10] use contrastive objective functions to pretrain speech models that are then finetuned. Alternatively, a common paradigm is to use a stronger teacher model to label unlabelled data [11] however this approach cannot be applied to the resource constrained setting of on-device learning. Noisy student learning or iterative pseudo-labelling approaches [12, 13] use the ASR model to self-label clean audio with the model trained to predict the same label with augmented version of the audio. Here the audio could be additionally filtered to include cases where the model does not have low confidence. We build off the work in [14] where hybrid HMM-DNN and connectionist temporal classification (CTC) ASR models are updated using a paired teacher model updated using an exponential moving average of the student model. These methods have not been applied to recurrent neural network-transducer (RNN-T) ASR models [15] that are streaming compatible and widely used across ASR applications.

We combine self-learning in this work with weak supervision. In conversational agents, users interact across multiple turns in a session. As shown in prior works [16], later interactions can be used to determine if a request has been correctly handled. If a user cancels or repeats their request, dissatisfaction is signalled. The semantics of the terminal request can be used as feedback for the initial request. Although this is not the ground truth transcription, we use such signals to update ASR models. Users can also be prompted for an explicit feedback signal as another example for a feedback score. We use the REINFORCE [17, 18, 19, 20] framework to update models using arbitrary rewards.

Contributions: We look at incremental updates to ASR models using unlabelled audio on edge devices with federated, compute and memory constraints. We show on public and internal datasets that:

- Self-learning with a paired teacher model updated through exponential moving average of ASR can be used to improve the performance of RNN-T by 10% on new use cases;
- Rehearsal training using historical datasets for generating model updates (pseudo-devices) at the cloud mitigates catastrophic-forgetting [21] on other test sets in self-training;
- Self-learning performance is improved by including weak supervision of NLU semantics or noisy feedback scores integrated through a policy-gradient approach.

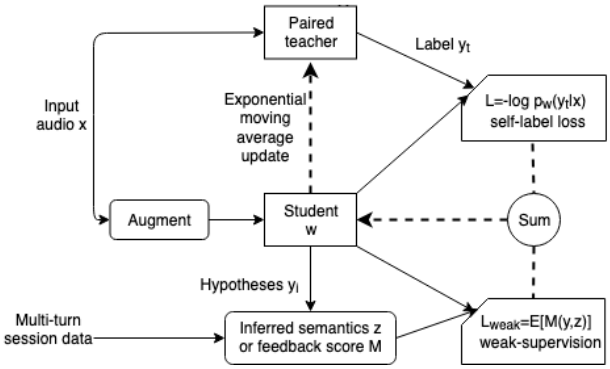

Fig. 1. Workflow of federated self-learning with weak supervision: a paired teacher model produces a label with clean audio. The self-label loss enforces consistency for the student with augmented audio. Alternate weak supervision loss minimizes the expected feedback score for a hypothesis inferred from multi-turn session data.

2. METHODS

2.1. RNN-T ASR model architecture

The RNN-T [15] architecture used for real-time speech recognition consists of a model that predicts the probability $P(\mathbf{y}|\mathbf{x})$ of labels $\mathbf{y} = (y_1, ..., y_U)$ given acoustic features $\mathbf{x} = (x_1, ..., x_T)$. It comprises an encoder, a prediction network, and a joint network. The encoder is analogous to an acoustic model that takes a sequence of acoustic input features, and outputs encoded hidden representations. The prediction network corresponds to a language model that accepts the previous output label predictions, and maps them to corresponding hidden representations. The joint network is a feed forward network that takes both the encoder and prediction network hidden representations, and predicts the final output label probabilities with softmax normalization. A model with parameters w produces m-hypotheses $(\mathbf{y}_1, ...\mathbf{y}_m)$ given input $\mathbf{x}$ with probability $p_w(\mathbf{y}_i|\mathbf{x})$.

2.2. Federated Self-Learning for ASR

Semi-supervised learning approaches typically employ a strong teacher model to machine transcribe audio data, which enables learning in the absence of human labeled supervised data. In compute, communication and memory constrained settings such as on-device federated learning, larger teacher models with higher resource requirements may not be feasible. In this work, we conform with *federated* constraints, and assume that the teacher model is of an equivalent configuration to the student model, can be stored and run on-device, and is used to process audio for machine labeling.

Algorithm 1 presents the details of the self-learning method. In each training round, we have unlabelled audio from the device for which we obtain the labels using the paired teacher model filtered to exclude utterances of very low or high confidence. Multiple local update steps may be taken on each device (similar to *FedAvg* [2]), or a single gradient update step may be taken (similar to *FedSGD*). The gradients are obtained using unlabeled audio on-device, with an augmented form of the audio and the teacher label. The server update step uses the aggregated local model deltas as a pseudo-gradient for its update. Finally, at the end of each training round based on an update frequency, the teacher model is updated using an exponential moving average (EMA [14]) of itself and the latest updated global student ASR model. This setup is illustrated in Fig. 1.

To help the model mitigate error feedback loops and catastrophic forgetting on older test sets, batches consisting of historical utterances with ground truth transcriptions can be included along with the self-learning updates that use unlabeled data. This process is termed as *rehearsal training*. The rehearsal updates are performed on the cloud by treating the cloud servers as a pseudo-devices and serves as a regularization term to prevent worsening ASR performance.

Algorithm 1 Federated self-learning for updating ASR with optional weak supervision including cloud pseudo-devices

Require: N number of local steps per round, η, η_k - global and on-device learning rates, δ exponential moving average rate (EMA), u EMA update frequency, $\mathcal{L}(\cdot)$ RNN-T loss function optionally including $\mathcal{L}_{weak}(\cdot)$ weak supervision loss if such weak feedback is available on-device, $\mathcal{D}_{ht}$ data used for rehearsal training with ground-truth labels, $\mathcal{C}$ cloud pseudo-devices.

Ensure: $w^r_{\mathcal{G}}$ incrementally updated global model, $w^r_{\mathcal{T}}$ updated teacher model

Init. $w^0_{\mathcal{G}}$ {start training with a pre-trained model}
for each round $r = 1, 2, \ldots$ **do**
  $\mathcal{S} \leftarrow$ (sample a subset of devices)
  for each device $k \in \mathcal{S} \cup \mathcal{C}$ **in parallel do**
    $w^r_k = w^{r-1}_{\mathcal{G}}$ {Global models broadcasted}
    if $k \in \mathcal{S}$ **then**
      $\mathcal{D}_{train} \leftarrow$ (draw from $\mathcal{D}_{ssl}$ filtered utterances on device k transcribed by teacher model $w^r_{\mathcal{T}}$, including weak supervision information z and audio augmentation)
    else if $k \in \mathcal{C}$ **then**
      $\mathcal{D}_{train} \leftarrow$ (draw from $\mathcal{D}_{ht}$ for ASR rehearsal training on cloud pseudo-devices)
    end if
    for batch $b_i, \forall i \in \{1, \ldots, N\}$ from $\mathcal{D}_{train}$ **do**
      $w^r_k \leftarrow \texttt{optimizer}_k\texttt{.update}(\eta_k, \nabla\mathcal{L}(w^r_k; b_i))$
    end for
    $\Delta w^r_k = w^r_k - w^{r-1}_{\mathcal{G}}$ {Transmit model delta to cloud}
  end for
  $w^r_{\mathcal{G}} \leftarrow \texttt{optimizer.update}(\eta, \frac{1}{|\mathcal{S}\cup\mathcal{C}|}\sum_{k\in\mathcal{S}\cup\mathcal{C}} \Delta w^r_k)$
  $w^r_{\mathcal{T}} \leftarrow \begin{cases} \delta w^{r-1}_{\mathcal{T}} + (1-\delta) w^r_{\mathcal{G}} \,, r \equiv_u 0 \\ w^{r-1}_{\mathcal{T}} \quad \text{else.} \end{cases}$ {EMA update}
end for

2.3. Weak supervision

Weak supervision signals can be used to further improve the performance of the system by leveraging information beyond just the unlabeled audio that self-learning relies on. This work exploits information weaker than the ground truth ASR transcription, which could be recovered from user interactions with the conversational agent. For example, if a user *stops*, *cancels* or *repeats* a request in the subsequent turn of a dialog, it indicates that the previous query was unsuccessfully processed by the device. We study updating ASR models with the help of such a feedback score, potentially indicating whether the user's request was unsuccessful. Further, the correct natural language understanding (NLU) semantics in the form of the correct slot value may eventually be recovered, for e.g., through an explicit re-invocation by the user. Hence, we also study leveraging weak feedback in the form of the NLU slot labels. An example of weak supervision for an utterance can be seen in Table 1.

In this work, we demonstrate the impact of weak supervision labels in two forms: (1) machine generated NLU semantics: from an alternate spoken language understanding (SLU) built from ASR→NLU as a proxy for inferred semantics from user session data; (2) synthetic user feedback scores: a proxy for real user corrections, and available only for the hypothesis served to the user. This framework can accommodate many types of weak supervision information.

Table 1. Examples of weak supervision available for an utterance. Here, semantic cost (fraction of slots incorrect) is illustrated as the feedback signal.

Transcription	play Halo by Beyonce in main speaker
ASR hypothesis	play Hello by Beyond in main speaker
NLU semantics	PlaySong, Artist:Beyonce, Song: Halo, Device: Main speaker
semantic cost	2/3

2.3.1. Weak Supervision: NLU semantics

Machine generated NLU semantics from an alternative ASR and NLU model are used as a form of weak NLU feedback, e.g. prior work [16] has used NLU feedback generated by rewriting utterances. Treating the NLU semantics z consisting of the slot type and values from this alternate system as ground truth, we can compute a semantic cost metric $M(z, \mathbf{y}_i)$ for an ASR hypothesis. The semantic cost metric is computed for a given hypothesis, as the fraction of slots that have an error. A slot is considered to have an error if the tokens within the slot are not all present in the hypothesis. For the purpose of experimentation, we also study the impact of using the alternate system's ASR transcript in addition to the NLU semantics. In this case, the cost M can include the word error rate (WER) obtained comparing $\mathbf{y}_i$ with the alternate transcript z_t. For ease of exposition, we consider z to encapsulate both semantics and transcription z_t.

To leverage feedback from these possibly erroneous NLU semantics, we train a model with weight w where the self-learning loss is augmented (summed) with this loss term from the weak NLU signal:

$$\begin{aligned}\mathcal{L}_{\text{weak}}(w, \mathbf{x}, z) &= \mathbb{E}_{y\sim p_w(\mathbf{y}|\mathbf{x})}[M(\mathbf{y}, z)] \\ &\approx \sum_i \hat{p}_w(\mathbf{y}_i|\mathbf{x}) M(y_i, z) \qquad (1) \\ \implies \nabla_w \mathcal{L}_{\text{weak}}(w, \mathbf{x}, z) &\approx \sum_i M(\mathbf{y}_i, z) \nabla_w \hat{p}_w(\mathbf{y}_i|\mathbf{x}),\end{aligned}$$

where $\hat{p}_w(\mathbf{y}_i|\mathbf{x}) = p_w(\mathbf{y}_i|\mathbf{x}) / \sum_j p_w(\mathbf{y}_j|\mathbf{x})$ is the normalized probability of the hypothesis. By making an assumption in (1), that the probability mass is concentrated in the n-best hypothesis of ASR, the expectation can be approximated by only considering this subset of hypotheses [20]. We note that p_w is a differentiable function of w and hence a gradient $\nabla_w \mathcal{L}$ can be computed.

2.3.2. Weak Supervision: Feedback Scores

In Sec. 2.3.1, we made an assumption that we can obtain weak NLU semantics, and thus have feedback for any hypothesis $\mathbf{y}_i$. Here, we add a constraint that weak supervision is only available for the hypothesis served to the user. The formulation with this constraint, termed weak supervision based on feedback scores, more closely simulate real user feedback where the user has provided feedback only for the served recognition.

We study two forms of feedback scores - (1) the semantic score as detailed in Sec. 2.3.1 applied only to the served hypothesis and (2) a binary feedback cost based on the sentence error rate with the true transcription z_t, $M(\mathbf{y}, z_t) = \mathbb{1}(\mathbf{y} \neq z_t)$ (as a proxy for binary user corrections). To simulate an estimation error of the feedback from user interactions, we add a noise term to the feedback signal obtained i.e. $M'(\mathbf{y}, z) = M(\mathbf{y}, z) + U$, with random variable U arising from an arbitrary noise distribution. This helps capture asymmetry and non-uniformity in the feedback from user interactions.

The learning is performed with a policy gradient setup. We use the n-best hypotheses to approximate the output lattice/space. A hypothesis (*action*) is selected from it by sampling based on the normalized n-best hypotheses probabilities. For the selected hypothesis, we use the feedback $M'(\mathbf{y}, z)$ described above as a *reward function* for the policy gradient method to update w which in turn parameterizes $\hat{p}_w(\mathbf{y}_i|\mathbf{x})$. We use the REINFORCE [17, 20] trick in conjunction with the above to obtain gradients so as to update w. Now,

$$\begin{aligned}\nabla_w \mathcal{L}_{\text{weak}}(w, \mathbf{x}, z) &= \mathbb{E}_{\mathbf{y}\sim p_w(\mathbf{y}|x)}[M(\mathbf{y}, z) \nabla_w \log(p_w(\mathbf{y}|\mathbf{x}))] \\ &\approx M'(\mathbf{y}, z) \nabla_w log(p_w(\mathbf{y}|\mathbf{x})), \mathbf{y} \sim p_w(\cdot|\mathbf{x}),\end{aligned}$$

where we take a sampling approximation of size 1 as an estimate of the expectation. With the above setup in place, this framework falls into the premise of Algorithm 1.

3. EXPERIMENTS

Data

Our federated continual training experiments are run from January to June 2021. We use an internal voice-assistant dataset with de-identified utterances totalling 4500 hours in this time period from 800K devices. We make only a *single pass* through this data as one of the constraints is that persistent audio storage is not feasible.

We evaluate the models on in-house human transcribed (HT) test sets. There is no speaker overlap between the train and evaluation datasets. *General* comprises a 37-hour test set in 2021 and older test sets in 2020. *Delta* comprises a 22-hour HT test set that records a change in frequency of words in 2021 over 2020. The transcriptions are filtered based on 1-gram, 2-gram and 3-grams that are 5x more frequent in 2021 than 2020. This test set captures changes in the data distribution such as new use cases and is crucial to measure the impact of continual learning.

We also demonstrate results on models trained on public test sets. We use RNN-T models pretrained on the 960 hour Librispeech dataset [22] and finetuned using self-learning with weak supervision on the 56 hour SLURP dataset [23]. For the public SLURP dataset, we evaluate on the test partition with 13K utterances.

Model

The RNN-T model used contains 60M parameters with a 5×1024 LSTM encoder, a 2×1024 LSTM prediction network and a feed-forward joint network with *tanh* activation [24]. The input embeddings of the prediction network are 512 dimensional. We use a 2500 sub-word piece tokenizer [25]. The audio features are 64 dimensional log-mel filter-bank energy features that are computed on a 25ms window, with a 10ms shift. SpecAugment [26] is used for the audio features. The features computed on 3 consecutive 10ms frames are stacked and sub-sampled to result in 192 dimensional features at a 30ms frame rate, provided as input to the ASR model.

A $480K$-hour pre-training dataset (where 120K hours are human transcribed and rest machine transcribed) is utilized for pre-training the baseline. Experiments using multiple losses, have equally weighted losses (no tuning). All results shown are using FedSGD with 400 devices randomly chosen for each of 3000 training rounds, batch size 16 and server-side Adam optimizer. For rehearsal training, 40 cloud pseudo-devices additionally used with historic transcribed data.

Metric

The performance of these models on the voice-assistant data is measured in terms of relative word error rate reduction (WERR) over the initial baseline model at the start of 2021. Positive WERR values represent improvements, while negative ones show degradations. Absolute WER numbers are reported on SLURP experiments.

Table 2. Performance of federated self-learning with weak supervision on the SLURP dataset, including examples of corrected utterances.

Setting	WER
Initial	28.70
Oracle supervised finetuning	16.95
Self-learning	
Teacher not updated	23.52
Teacher updated with EMA	18.95
+weak-supervision	18.79

Truth	please help me turn on the robot vacuum cleaner
Initial	please tell me turn on the roblox i can clean
Self-learn	please tell me turn on the robot vacuum cleaner
Truth	look for this playback in audiobook and play for me
Initial	look for display light audiobook and play for me
Self-learn	look for this playback in audiobook and play for me
Truth	olly what else do i have on the list
Initial	what else do i have in the list
Self-learn	ollie what else do i have on the list

Table 3. Performance of federated self-learning with weak supervision on voice-assistant data. WERR numbers are relative to WER of the initial model. Multiple forms of weak supervision such as ASR and NLU labels from an alternate SLU model, and NLU feedback scores for the hypothesis served are contrasted.

Weak Supervision method	Teacher Update	General WERR	Delta WERR
-	-	-8.16	-0.02
-	✓	-6.12	8.29
ASR	✓	-1.84	11.43
ASR + NLU	✓	-1.22	11.56
NLU feedback-score	✓	-1.64	12.06

4. RESULTS

Federated self-learning with weak supervision: We see the performance of self-learning of a pretrained RNN-T model on the public SLURP dataset in Table 2 that shows self-learning improving the performance by 19% with additional gains from using weak supervision composed of NLU feedback scores. We note that limited gains arise from weak supervision as SLURP has sparse annotations for transcript tokens or few slots per utterance. In few corrected examples, we see self-learning with weak supervision correcting deletion errors and even learning new words like the keyword 'olly'.

In Table 3, the performance of self-learning coupled with weak supervision is depicted for continual learning with a single pass on the internal dataset. First, we observe that if we do not update the paired teacher model with EMA, performance on the new use case does not improve. If we only do self-learning for ASR, there is an improvement of 8.3% on the new use case test set. Coupling this with an ASR based weak supervision (where each hypothesis gets a feedback score of the WER computed using a teacher model), we see more improvement that increases as feedback includes the NLU component. We also see similar improvement using only the NLU-based feedback-score obtained only for the served hypothesis as opposed to obtaining a score for all possible hypotheses.

Noisy feedback: Table 4 shows the result of federated learning only with noisy feedback for a single served hypothesis from ASR. Here we consider noisy feedback of the form, $M'(\mathbf{y}, z) = M(\mathbf{y}, z) + (-1)^{M(\mathbf{y},z)}U'$, where random variable $U' \sim p(U|U \in [0, 1]), U \sim \mathcal{N}(0, \sigma^2)$ is drawn from a normal random variable with variance σ^2 truncated to be in the range $[0, 1]$. We then add different levels of noise to measure its impact. In a noisy version of a binary feedback score,

$$\begin{aligned} \mathbb{E}\left[M'(\mathbf{y}, z)\right] &= \mathbb{E}[M(\mathbf{y}, z)] + \mu\,\mathbb{E}[(-1)^{M(\mathbf{y},z)}] \\ &= (1 - 2\mu)\,\mathbb{E}[M(\mathbf{y}, z)] + \mu \\ \implies \nabla_w\,\mathbb{E}[M'(\mathbf{y}, z)] &= (1 - 2\mu)\nabla_w\,\mathbb{E}[M(\mathbf{y}, z)], \end{aligned}$$

where $\mu = \mathrm{E}[U']$. Thus if the mean is less than 0.5, gradient update with the noisy feedback, in expectation, is in the same direction as the gradient update with the true feedback. We demonstrate that even at a high level of noise of $\sigma = 0.4$ we are still able to improve the model on the delta dataset significantly.

Table 4. Performance of learning with only noisy feedback scores on voice-assistant data

Setting	Delta WERR
binary feedback without noise	14.45
binary feedback + noise ($\sigma = 0.1$)	9.05
binary feedback + noise ($\sigma = 0.2$)	7.41
binary feedback + noise ($\sigma = 0.4$)	4.40

Table 5. We study (i) the effect of rehearsal training in mitigating the catastrophic forgetting (left) and (ii) the effect of hyper parameters (right) in self-learning on voice-assistant data

Setting	Delta WERR	General (2020) WERR
Self-learning	14.08	-13.63
+ rehearsal training	12.47	-5.85

ema δ, *update* u	Delta WERR
0.999, 10	**14.08**
0.999, 100	10.38
0.999, 200	11.56
0.9999, 10	12.64
0.9999, 100	11.03
0.975, 1	diverge

EMA hyperparameters and rehearsal training: In Table 5, we first see the impact of rehearsal training on mitigating catastrophic forgetting - we observe reduced regression on the older 2020 test set at the expense of performance of new *Delta* test sets. Delta test set results are not comparable across prior tables as amount of computation, catastrophic forgetting differ. We also study the impact of EMA hyperparameters, higher δ implies lower weight to new updates and update frequency u determines how often the teacher model is updated. Improved performance is seen for frequent updates with a lower EMA value. We also observed training diverging when the teacher model is updated to the student model after each step, suggesting that an error feedback loop takes place.

5. CONCLUSION

We focused on the federated continual learning problem for ASR where an ASR model deployed on-device is updated ensuring that (1) human ground-truth transcriptions are not available, (2) large device compute and memory are not required to run strong teacher models for labelling the audio (3) audio is not persisted or sent to the cloud. We demonstrated that using a paired teacher model to generate labels for the unlabelled audio and where the teacher model is updated using an exponential moving average of the RNN-T model can improve RNN-T performance by 10% on new use cases with larger improvement on public SLURP dataset and only 10% away from the fully supervised setting. Rehearsal training using historical datasets with ground-truth transcriptions mitigates catastrophic forgetting and error feedback loops. We made use of weak supervision signals such as machine generated NLU semantics or simulated noisy feedback scores from interactions of a user in a policy-gradient approach which further improved the performance of self-learning.

Acknowledgments: We thank Gurpreet, Aaron, Buddha, Bach, Harish, Ehry and Shehzad for helpful discussions.

6. REFERENCES

[1] M. Al-Rubaie and J. M. Chang, "Privacy-preserving machine learning: Threats and solutions," *IEEE Security & Privacy*, vol. 17, no. 2, pp. 49–58, 2019.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[3] D. Guliani, F. Beaufays, and G. Motta, "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.

[4] A. Hard, K. Partridge, C. Nguyen, N. Subrahmanya, A. Shah, P. Zhu, I. L. Moreno, and R. Mathews, "Training keyword spotting models on non-iid data with federated learning," *arXiv preprint arXiv:2005.10406*, 2020.

[5] F. Granqvist, M. Seigel, R. van Dalen, Á. Cahill, S. Shum, and M. Paulik, "Improving on-device speaker verification using federated learning with privacy," *arXiv preprint arXiv:2008.02651*, 2020.

[6] A. Hard, K. Partridge, N. Chen, S. Augenstein, A. Shah, H. J. Park, A. Park, S. Ng, J. Nguyen, I. L. Moreno *et al.*, "Production federated keyword spotting via distillation, filtering, and joint federated-centralized training," *arXiv preprint arXiv:2204.06322*, 2022.

[7] Z. Huo, D. Hwang, K. C. Sim, S. Garg, A. Misra, N. Siddhartha, T. Strohman, and F. Beaufays, "Incremental layer-wise self-supervised learning for efficient unsupervised speech domain adaptation on device," *Proc. Interspeech 2022*, pp. 4845–4849, 2022.

[8] J. Jia, J. Mahadeokar, W. Zheng, Y. Shangguan, O. Kalinli, and F. Seide, "Federated domain adaptation for asr with full self-supervision," *arXiv preprint arXiv:2203.15966*, 2022.

[9] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv preprint arXiv:2202.03555*, 2022.

[10] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[11] S. H. K. Parthasarathi and N. Strom, "Lessons from building acoustic models with a million hours of speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6670–6674.

[12] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau, R. Collobert, G. Synnaeve, and M. Auli, "Self-training and pre-training are complementary for speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3030–3034.

[13] Y. Chen, W. Wang, and C. Wang, "Semi-supervised asr by end-to-end self-training," *arXiv preprint arXiv:2001.09128*, 2020.

[14] V. Manohar, T. Likhomanenko, Q. Xu, W.-N. Hsu, R. Collobert, Y. Saraf, G. Zweig, and A. Mohamed, "Kaizen: Continuously improving teacher using exponential moving average for semi-supervised speech recognition," *arXiv preprint arXiv:2106.07759*, 2021.

[15] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[16] P. Ponnusamy, A. R. Ghias, C. Guo, and R. Sarikaya, "Feedback-based self-learning in large-scale conversational ai agents," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 08, 2020, pp. 13 180–13 187.

[17] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[18] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *Interspeech*, vol. 2013, 2013, pp. 2345–2349.

[19] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C.-C. Chiu, and A. Kannan, "Minimum word error rate training for attention-based sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4839–4843.

[20] M. Rao, P. Dheram, G. Tiwari, A. Raju, J. Droppo, A. Rastrow, and A. Stolcke, "Do as i mean, not as i say: Sequence loss training for spoken language understanding," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7473–7477.

[21] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[23] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "Slurp: A spoken language understanding resource package," *arXiv preprint arXiv:2011.13205*, 2020.

[24] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[25] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *ACL*, 2018.

[26] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.