Federated Acoustic Model Optimization for Automatic Speech Recognition

Conghui Tan[1(✉)], Di Jiang[1], Huaxiao Mo[1], Jinhua Peng[1], Yongxin Tong[2], Weiwei Zhao[1], Chaotao Chen[1], Rongzhong Lian[1], Yuanfeng Song[1], and Qian Xu[1]

[1] AI Group, WeBank Co., Ltd., Shenzhen, China
{martintan,dijiang,vincentmo,kinvapeng,davezhao,chaotaochen, ronlian,yfsong,qianxu}@webank.com
[2] BDBC, SKLSDE Lab and IRI, Beihang University, Beijing, China
yxtong@buaa.edu.cn

Abstract. Traditional Automatic Speech Recognition (ASR) systems are usually trained with speech records centralized on the ASR vendor's machines. However, with data regulations such as General Data Protection Regulation (GDPR) coming into force, sensitive data such as speech records are not allowed to be utilized in such a centralized approach anymore. In this demonstration, we propose and show the method of federated acoustic model optimization in order to solve this problem. This demonstration does not only vividly show the underlying working mechanisms of the proposed method but also provides an interface for the user to customize its hyperparameters. With this demonstration, the audience can experience the effect of federated learning in an interactive fashion and we wish this demonstration would inspire more research on GDPR-compliant ASR technologies.

Keywords: Automatic Speech Recognition · Federated learning

1 Introduction

Automatic Speech Recognition (ASR) is becoming the premise of a variety of modern intelligent equipments. The performance of contemporary ASR systems heavily rely on the robustness of acoustic models, which are conventionally trained from speech records collected from diverse client scenarios in a centralized approach. However, due to the increasing awareness of data privacy protection and the trends of strict data regulation such as the European Union General Data Protection Regulation (GDPR) taking effect, collecting clients' speech data and utilizing them in a centralized approach is becoming prohibited. Therefore, in the new era of strict data privacy regulations, a new paradigm is heavily needed to make it possible for the ASR vendors to consistently train or refine their acoustic models.

The video of this paper can be found in https://youtu.be/H29PUN-xFxM.

Y. Nah et al. (Eds.): DASFAA 2020, LNCS 12114, pp. 771–774, 2020.
https://doi.org/10.1007/978-3-030-59419-0_54

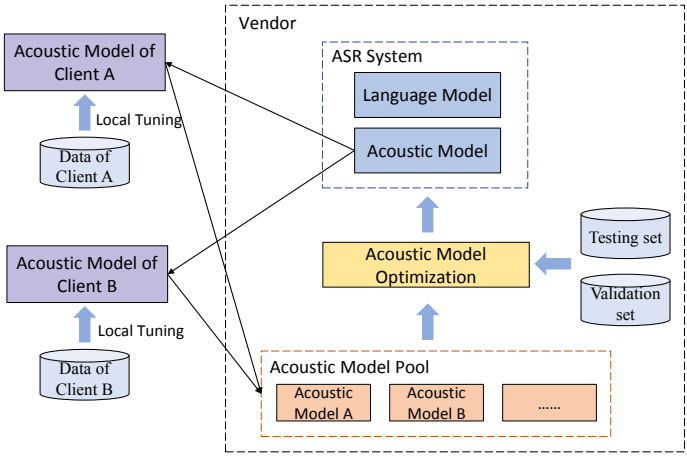

Fig. 1. System architecture

In this demonstration, we propose and demonstrate a novel paradigm of optimizing acoustic model for ASR vendors based on federated learning [2,6], which is a distributed training framework and enables the clients to collaboratively train a machine learning model without sharing the their private data. Through this demonstration, the audience would see that our proposed technique provides an effective approach for ASR vendors to refine their acoustic model in a privacy-preserving way and it potentially lay the foundation of developing more GDPR-compliant technologies for the ASR industry.

2 System Architecture

As shown in Fig. 1, our system consists of one vendor's central server and multiple clients. Private speech data is kept on the clients. The server maintains a global acoustic model, while each client has its local acoustic model, and server and clients will collaboratively optimize their models under the federated learning framework.

In each round of training, vendor will first distribute its acoustic model to all the clients, and each client optimize the received model on its local data. After summarizing the changes in the local model, only these changes are sent back to the vendor's server by using encrypted communication. After that, servers collects the updates from the clients, and tries to these updates to obtain a better global acoustic model. During the whole process, all the training data always stays on the clients' devices, and there is no worry for clients about leaking their local sensitive data to the vendor or the third-party.

Tuning on Client Data. To optimize acoustic model on the local data, transfer learning technique is adopted. More specially, we use model-based transfer learning with KLD-regularization [1], which aims at training a new model with

lower training loss on the local data while keeping the KL divergence between the trained model and the vendor's original model not too large. In this way, we can obtain a new local model which is more adapted to the client's local setting but also robust to other scenarios. For the language model, it is also tuned on each client via transfer learning, but we will not synchronize it with the vendor via federated learning.

Acoustic Model Optimization. The tuned local models contains information of the extra local training data, and thus can be naturally be used for improving the global model. As a result, a new issue comes: how can we combine the updated models collected from different clients? Directly averaging them is a natural choice, but not good enough. Here we propose to merge the models by genetic algorithm [3]. In genetic algorithm, we need to maintain a set of candidate chromosomes, which are actually the encoded expressions for the models, and include all the collected models at initial stage. In each iteration of genetic algorithm, we generate new candidates by randomly mutating, crossovering and weighted-averaging the chromosomes in the candidate set. After that, we evaluate the performance of each candidate model by calculating its word error rate (WER) on the validation set, then the models with poor performances are eliminated from the set. After repeating such iterations for a certain time, we stop the algorithm and choose the candidate with lowest WER as the new global model.

3 System Implementation

The ASR system testbed is built upon the open source ASR toolkit Kaldi [4]. Without loss of generality, we utilize the Kaldi "chain" model as the acoustic model and the backoff n-gram model as the language model. The backoff n-gram language model is trained through the SRILM toolkit [5]. The whole system is deployed on multiple machines, one of which is the vendor node and the others are the client nodes. The hardware configuration of each machine is 314 GB memory, 72 Intel Core Processor (Xeon), Tesla K80 GPU and CentOS.

4 Performance Evaluation

We proceed to show a quantitative evaluation result of the proposed technique in a scenario of ten clients. The original vendor ASR system is trained with 10,000 h of speech data and each client has about 100 h of private speech data. Another 100 h of speech data is reserved for validation and testing. We compare the performances of the vendor's original acoustic model, the acoustic models tuned on clients' data, and the final optimized acoustic model of the vendor in terms of Word Error Rate (WER) on testing data. The experimental result is shown in Fig. 2. We observe that the optimized acoustic model significantly outperforms its courtparts, showing that the proposed technique can enable the ASR vendors to concisely refine its ASR system in the new era of strict data privacy regulations. In our demonstration, the audience is able to freely configure the experimental setting and observe the effects of changing the number of clients or their dataset.

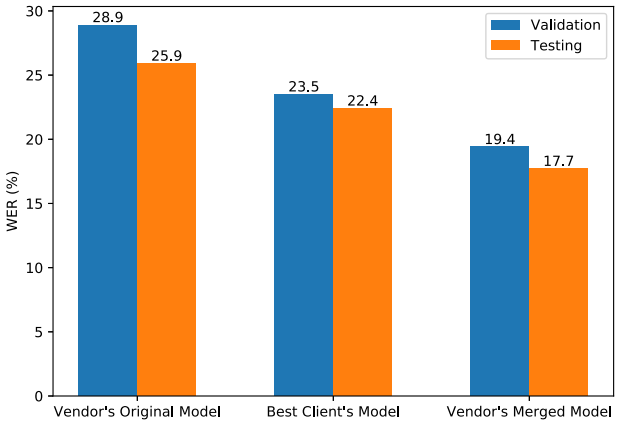

Fig. 2. Word Error Rates (WER) of each model

5 Conclusion

In this demonstration, we show a novel technique of federated acoustic model optimization, which solves the most critical problem faced by the ASR industry in the new era of strict data regulations. Through this demonstration, the audience will have a unique opportunity of experiencing how federated learning protects the clients' data privacy while provides the vendors sufficient information to further refine their acoustic models. We wish that the demonstration would shed light on developing better GDPR-compliant ASR technologies.

References

1. Huang, Y., Yu, D., Liu, C., Gong, Y.: Multi-accent deep neural network acoustic model with accent-specific top layer using the KLD-regularized model adaptation. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
2. Konečnỳ, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D.: Federated learning: strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492 (2016)
3. Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press, Cambridge (1998)
4. Povey, D., et al.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (2011)
5. Stolcke, A.: SRILM-an extensible language modeling toolkit. In: Seventh International Conference on Spoken Language Processing (2002)
6. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. (TIST) **10**(2), 12 (2019)