A Platform for Deploying the TFE Ecosystem of Automatic Speech Recognition

Yuanfeng Song
The Hong Kong University of Science and Technology & WeBank Co., Ltd
Hong Kong, China
songyf@cse.ust.hk

Rongzhong Lian
Yixin Chen
Di Jiang
Xuefang Zhao
Conghui Tan
Qian Xu
WeBank Co., Ltd
Shenzhen, China

Raymond Chi-Wing Wong
The Hong Kong University of Science and Technology
Hong Kong, China
raywong@cse.ust.hk

ABSTRACT

Since data regulations such as the European Union's General Data Protection Regulation (GDPR) have taken effect, the traditional two-step Automatic Speech Recognition (ASR) optimization strategy (i.e., training a "one-size-fits-all" model with vendor's centralized data and fine-tuning the model with clients' private data) has become infeasible. To meet these privacy requirements, TFE, a novel GDPR-compliant ASR ecosystem, has been proposed by us to incorporate transfer learning, federated learning, and evolutionary learning towards effective ASR model optimization. In this demonstration, we further design and implement a novel platform to promote the deployment and applicability of TFE. Our proposed platform allows enterprises to easily conduct the ASR optimization task using TFE across organizations.

CCS CONCEPTS

• **Computing methodologies** → **Speech recognition**;

KEYWORDS

federated learning, evolutionary learning, speech recognition

ACM Reference Format:
Yuanfeng Song, Rongzhong Lian, Yixin Chen, Di Jiang, Xuefang Zhao, Conghui Tan, Qian Xu, and Raymond Chi-Wing Wong. 2022. A Platform for Deploying the TFE Ecosystem of Automatic Speech Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3503161.3547731

1 INTRODUCTION

The global speech and Automatic Speech Recognition (ASR) software market is still growing steadily [1], due to its wide range of applications in industry [2, 10–12]. In existing ASR market, traditional ASR vendors usually provide a "one-size-fits-all" system trained with centralized speech data and then fine-tune the model with clients' private data. However, such a privacy-violating strategy has become infeasible since data regulations such as the European Union's General Data Protection Regulation (GDPR) [15] have taken effect and it is prohibited to access the clients' private speech data. Thus, a GDPR-compliant ASR ecosystem called **TFE** has been proposed by us to alleviate the aforementioned privacy concerns [5, 13, 14]. As shown in Fig. 1, TFE incorporates **T**ransfer learning, **F**ederated learning, and **E**volutionary learning towards effective privacy-preserving ASR model optimization across the involved parties (i.e., the vendor and the clients).

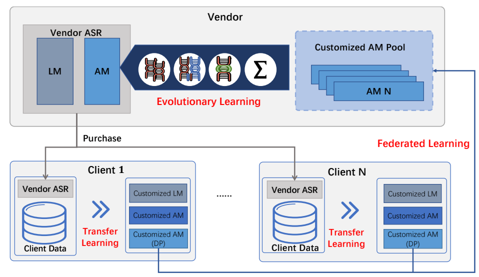


Figure 1: The workflow of our proposed TFE ecosystem

The operation of the TFE ecosystem requires seamlessly coordination of multiple parties, since the vendor and the clients are not necessary to be in the same organization. To save the manual labor involved and further increase the efficiency in deploying and maintaining TFE, in this demonstration, we present a novel platform to automatic TFE in ASR industry. More specifically, this platform is designed to use software programs as robots to automate the whole optimization process across the involved parties displayed in Fig. 1. Through this demonstration, users will experience the working mechanisms of the TFE ecosystem and how this platform can improve its efficiency. While there are already a substantial amount of federated learning studies in the research community, limited of them have explored the feasibility of deploying them into industrial scenarios, and we hope our proposed platform would shed new light on the deployment of more federated learning applications.

2 THE TFE ECOSYSTEM

Despite the increasing popularity of the end-to-end ASR models in the research community [3], the *hybrid* ASR systems still dominate the industry due to their flexibility and modularization [7]. Hence, in this demonstration, we mainly focus on the optimization of hybrid ASR systems. A hybrid ASR system is typically composed of an

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3547731

Acoustic Model (AM) and a Language Model (LM), where the AM is a deep neural network (DNN) or a DNN-HMM [8] responsible for translating the speech features into their phoneme representation, and the LM estimates the probability of the word sequences. As described in Fig. 1, TFE is composed of three machine learning components: transfer learning (TL), federated learning (FL), and evolutionary learning (EL).

TL for Client Given the vendor's "one-size-fits-all" model, TFE resorts to transfer learning to tune a highly customized ASR system for each client. Specifically, the customized LM is achieved by interpolating the vendor's LM with the client's LM (i.e., LM trained on client's private data) using

$$P_{CLM}(\mathbf{w}) = \lambda P_{LM}^{V}(\mathbf{w}) + (1-\lambda) P_{LM}^{C}(\mathbf{w}), \tag{1}$$

where $P_{CLM}(\mathbf{w})$ denotes the probability of the word sequence $\mathbf{w}$ given by the customized LM, $P_{LM}^{V}(\mathbf{w})$ and $P_{LM}^{C}(\mathbf{w})$ are probabilities delivered by the vendor's LM and the client's LM, respectively. For the AM, TFE adopts conservative training to conduct the model-based transfer learning for the DNN part [17]. The neural network loss of the target domain $\mathcal{L}_T$ can be transferred from the source domain loss $\mathcal{L}_S$ as

$$\mathcal{L}_T = (1-\eta)\mathcal{L}_S + \frac{\eta}{N} \sum_{(x,y)\in \text{Target Domain}} p_S(y|x) \log p_T(y|x), \tag{2}$$

where (x, y) denotes a data sample from the target domain, N represents the total number of data samples, η is a transfer ratio hyper-parameter, and $p_S(\cdot)$ and $p_D(\cdot)$ are the posteriors given by the source and the target domain model, respectively. Finally, the client will get a highly-customized ASR model for their domain.

FL between Client and Vendor TFE utilizes federated learning [16] to achieve privacy-preserving transmission of the customized AMs from each client. We propose a Differential Privacy AM (DP-AM) which encrypts the gradients of the mini-batch stochastic gradient descent (SGD) of DNNs by

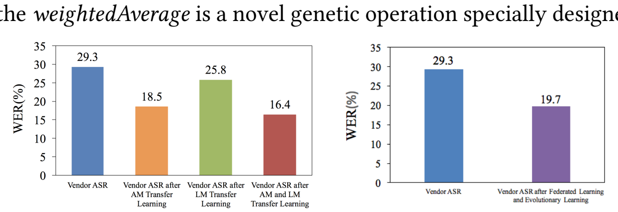
$$w_{t+1} = w_t - \eta_t(\Delta w_t' + \delta_t), \tag{3}$$

where η_t is the learning rate, δ_t is a generated noise, and $\Delta w_t'$ is a scaled gradient given by

$$\Delta w_t' = \frac{\Delta w_t}{\max\{1, \ \|\Delta w_t\|_2/S\}}, \tag{4}$$

where Δw_t is the gradient, and S is the sensitivity in DP [4]. The federated learning component in TFE achieves privacy-preserving information collection for the vendor to improve its ASR system.

EL for Vendor Evolutionary learning is used in TFE to incorporate the customized AMs and renovate the vendor AM. We propose an Acoustic Genetic Algorithm (AGA) which involves four steps: *initialization, selection, genetic operators, and termination.* During each iteration, the AGA generates AM candidates using genetic operations, namely reproduction, crossover, mutation, and weightedAverage, and then selects the AM with the lowest word error rate (WER) [6] as the new vendor AM. Inspired by FedAvg algorithm [9], the *weightedAverage* is a novel genetic operation specially designed



(a) TL for Client Side (b) FL & EL for Vendor Side

Figure 2: The WER comparison by the TFE ecosystem

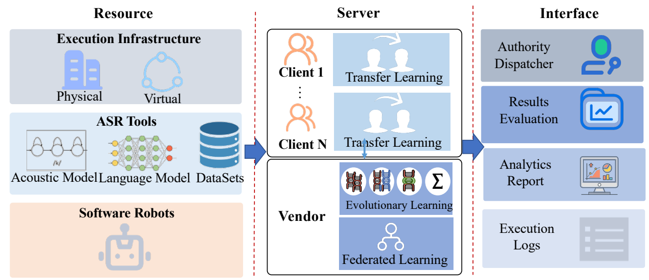


Figure 3: The system architecture of our proposed platform

by us for ASR model optimization [5], and it creates the parameters of a child by weighted averaging the parameters of its two parents.

TFE Performance Analysis The TFE is built upon the open-sourced Kaldi toolkit, and deployed on a 11-node cluster. Around 10,000 hours of speech data is used to train the vendor's ASR model, and another 250 hours data are used as the testing dataset. Each client is provided with diverse domains data ranging from radio programs to daily conversations as private data. Experimental results in Fig. 2 show that TFE can significantly improve the ASR performance (i.e., WER) for both local clients and vendor.

3 THE PROPOSED PLATFORM

The objective of our proposed platform is to provide a flexible tool to automate the TFE tasks across organizations.

System Architecture As shown in Fig. 3, the system roughly contains three layers: the Resource layer, the Server layer, and the Interface layer. The Resource layer servers as the foundation to run the whole ecosystem, and it includes typical components such as the infrastructure and the tools to construct the ASR models. The Server layer includes the main implementation of the transfer, federated, and evolutionary learning components. The Interface layer enables the user to check the statistics of the TFE ecosystem.

User Interface and Functionalities Fig. 4 shows the user interface (UI) of our proposed platform, including the following parts.

TL To conduct transfer learning for the local ASR model, users can upload their client's LM and AM or their private speech data. The customized LM and AM will be delivered with defined parameters.

FL After the users in the client side configure their local model or data path, TFE will conduct federated learning to transmit data-preserving mini-batch gradient to the vendor side. A detailed version history of the transmitted models is also provided for reference.

EL Users can choose the initial models as parents, and configure the number of generations. After the learning converges, the system will show the WER reduction and the updated vendor AM.

Model Evaluation This functionality evaluates the effectiveness of generated models in the model library with user-defined data.

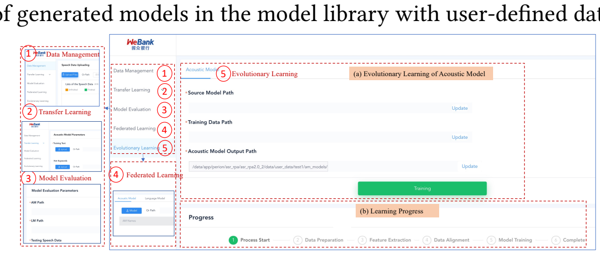


Figure 4: A UI screenshot of our proposed platform

Acknowledgement. The research of Yuanfeng Song and Raymond Chi-Wing Wong is supported by PRP/026/21FX.

REFERENCES

[1] Last accessed 13 May. 2022. Automatic Speech Recognition (ASR) Software Market is expected to reach a substantially Growth by 2027. https://www.digitaljournal.com/pr/automatic-speech-recognition-asr-software-market-is-expected-to-reach-a-substantially-growth-by-2027-2

[2] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2020. Voxento: A Prototype Voice-controlled Interactive Search Engine for Lifelogs. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*. 77–81.

[3] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4960–4964.

[4] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.

[5] Di Jiang, Conghui Tan, Jinhua Peng, Chaotao Chen, Xueyang Wu, Weiwei Zhao, Yuanfeng Song, Yongxin Tong, Chang Liu, Qian Xu, et al. 2021. A GDPR-compliant Ecosystem for Speech Recognition with Transfer, Federated, and Evolutionary Learning. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 3 (2021), 1–19.

[6] Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication* 38, 1-2 (2002), 19–28.

[7] Jinyu Li, Rui Zhao, Eric Sun, Jeremy HM Wong, Amit Das, Zhong Meng, and Yifan Gong. 2020. High-Accuracy and Low-Latency Speech Recognition with Two-Head Contextual Layer Trajectory LSTM Model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7699–7703.

[8] Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin, and Hichem Sahli. 2013. Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 312–317.

[9] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 1273–1282.

[10] Zhenhui Peng, Kaixiang Mo, Xiaogang Zhu, Junlin Chen, Zhijun Chen, Qian Xu, and Xiaojuan Ma. 2020. Understanding User Perceptions of Robot's Delay, Voice Quality-Speed Trade-off and GUI during Conversation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[11] Vraj Shah, Side Li, Arun Kumar, and Lawrence Saul. 2020. SpeakQL: towards speech-driven multimodal querying of structured data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2363–2374.

[12] Yuanfeng Song, Raymond Chi-Wing Wong, Xuefang Zhao, and Di Jiang. 2022. VoiceQuerySystem: A Voice-driven Database Querying System Using Natural Language Questions. In *Proceedings of the 2022 International Conference on Management of Data*. 2385–2388.

[13] Conghui Tan, Di Jiang, Huaxiao Mo, Jinhua Peng, Chaotao Chen, Rongzhong Lian, Yuanfeng Song, Qian Xu, and Qiang Yang. 2020. Federated Acoustic Model Optimization for Automatic Speech Recognition. In *International Conference on Database Systems for Advanced Applications*. Springer.

[14] Conghui Tan, Di Jiang, Jinhua Peng, Xueyang Wu, Qian Xu, and Qiang Yang. 2020. A De Novo Divide-and-Merge Paradigm for Acoustic Model Optimization in Automatic Speech Recognition. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3709–3715.

[15] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* (2017).

[16] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–19.

[17] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in neural information processing systems*. 3320–3328.