AN EXPERIMENTAL STUDY ON PRIVATE AGGREGATION OF TEACHER ENSEMBLE LEARNING FOR END-TO-END SPEECH RECOGNITION

Chao-Han Huck Yang[1,2*] *I-Fan Chen*[2] *Andreas Stolcke*[2]
Sabato Marco Siniscalchi[1,3] *Chin-Hui Lee*[1]

[1]Georgia Institute of Technology, USA and [2]Amazon Alexa AI, USA
[3]Department of Electronic Systems, NTNU, Trondheim, Norway

ABSTRACT

Differential privacy (DP) is one data protection avenue to safeguard user information used for training deep models by imposing noisy distortion on privacy data. Such a noise perturbation often results in a severe performance degradation in automatic speech recognition (ASR) in order to meet a privacy budget ε. Private aggregation of teacher ensemble (PATE) utilizes ensemble probabilities to improve ASR accuracy when dealing with the noise effects controlled by small values of ε. We extend PATE learning to work with dynamic patterns, namely speech utterances, and perform a first experimental demonstration that it prevents acoustic data leakage in ASR training. We evaluate three end-to-end deep models, including LAS, hybrid CTC/attention, and RNN transducer, on the open-source LibriSpeech and TIMIT corpora. PATE learning-enhanced ASR models outperform the benchmark DP-SGD mechanisms, especially under strict DP budgets, giving relative word error rate reductions between 26.2% and 27.5% for an RNN transducer model evaluated with LibriSpeech. We also introduce a DP-preserving ASR solution for pretraining on public speech corpora.

Index Terms— privacy-preserving learning, automatic speech recognition, teacher-student learning, ensemble training.

1. INTRODUCTION

Automatic speech recognition (ASR) [1] is widely used in spoken language processing applications, such as smart device control [2], intelligent human-machine dialogue [3], and spoken language understanding [4]. To build ASR systems, a large collection of user speech [5] is often required for training high-performance acoustic models. Protecting information privacy and measuring numerical privacy loss, such as whether data from a specific user is used for model training, are becoming critical and prominent research topics for on-device speech applications [6, 7, 8, 9, 10, 11, 12].

Differential privacy (DP) introduces a noise addition scheme for information protection characterized by **measurable privacy budgets**. The noise level is defined by a privacy budget (e.g., controlled by a minimum ε value) in ε-DP [13, 14] for a given data set. Machine learning frameworks based on ε-DP have been shown effective against leakage of training data (e.g., human faces), model inversion attacks [15] (MIA) and membership inference, in which a query-free algorithm is used to generate highly confident test data similar to its training set. Deploying ε-DP in speech applications based on deep models has recently attracted much interest [16]. However, directly applying ε-DP perturbation on the training data could lead to severe performance (e.g., prediction accuracy) degradation [17]. Therefore, the noise-enabled protection process needs to be designed carefully for incorporation into machine learning for training speech models. Most published DP-based approaches are still limited to recognition of isolated spoken commands [7]. Designing a continuous speech recognition system with ε-DP protection needs further investigation under a variety of privacy settings.

In this work we use ε-DP protection to show that acoustic features used in ASR training data can been protected against model inversion attack (MIA) [15]. We also demonstrate that ε-DP can prevent data leakage from a pretrained ASR model.

Private aggregation of teacher ensembles (PATE) [18] learning is a recently proposed framework that aims to avoid accuracy loss in large-scale visual prediction models while guaranteeing ε-DP protection. PATE is guided by a teacher-student learning process, where multiple teachers make up an ensemble for knowledge transfer. The idea underlying PATE is to benefit from aggregated noisy outputs of teacher models to compartmentalize nonsensitive public data with ε-DP protections. PATE and related approaches [19] were proven effective in mitigating model accuracy deterioration by employing an ensemble of teacher models with independent noise perturbations. Nonetheless, for ASR systems dealing with continuous speech, label stream prediction (e.g., by sequence level modeling) is required. In this study, we apply the PATE framework to continuous speech recognition and design different ensemble strategies to ensure ε-DP for

*Work done at Georgia Tech and Amazon. Parts of this study were completed while the first author was an intern at Amazon.

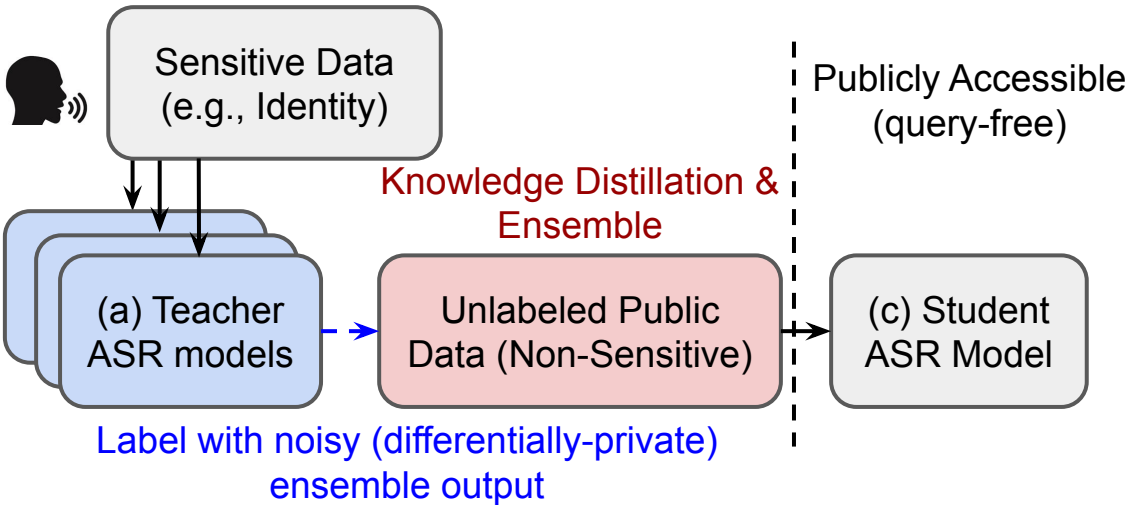


Fig. 1: Proposed framework for utilizing private aggregation teacher ensemble (PATE) to train end-to-end ASR with ε-differential privacy. Teacher ASR models could also be combined with pretraining on public data (our second case study).

end-to-end ASR models, including RNN transducer [20], hybrid CTC/attention [21], and LAS (Listener, Attender, Speller) [22] networks, as shown in the blue squares in Figure 1. Under a strict DP budget (ε=1), the PATE-trained end-to-end models maintain data privacy at the expense of only a slight increase in word error rate (WER). It also shows competitive advantages when compared to differentially private stochastic gradient descent (DP-SGD) [16, 23] benchmarks.

2. RELATED WORK

Recent research efforts to preserve data privacy in an ASR system can be categorized into two groups: (i) systemic, such as federated learning [24], data isolation [25], and data encryption [26], and (ii) algorithmic, mainly differentially private machine learning [16]. In this section, we first summarize some of the privacy-preserving solutions proposed for ASR. Next, we describe the substratum of differential privacy devised for machine learning applications and discuss their difference to our proposed approach, while highlighting its key contributions.

2.1. Privacy-Preserving Automatic Speech Recognition

Federated machine learning algorithms [24, 27, 25] have been investigated in the ASR community to improve information protection. For instance, the average gradient method [27] was deployed to update the model in ASR training. Vertical federated learning methods [25] show some other benefits from isolated features extractors under a heterogeneous computation. However, these system-level frameworks often rely on assumptions about the constrained accessibility of the malicious threats and barely provide universal and statistical justification for privacy guarantees. Cryptographic encryption [26, 28] and computation protocols [29] are established techniques for privacy-preserving speaker identification. Meanwhile, these encryption algorithms and protocols do not consider privacy protection for training samples, which is a crucial element in developing machine learning models on a large scale. Lately, differentially-private stochastic gradient descent (DP-SGD) [16, 23] was introduced to allow quantitative measurements and further protect identity inference (e.g., of accent or speaking condition) by introducing an additive distortion during model training. In the remainder of the paper, a mathematical formulation of differential privacy is provided, as well as a study of its effectiveness for ASR.

2.2. Differential Privacy Fundamentals

The differential privacy mechanism [13] is a standard method to deploy algorithms with a target privacy guarantee.

Definition 1. A randomized algorithm $\mathcal{M}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ is (ε, δ)-differentially private if for any two neighboring inputs (e.g., speech data) $d, d' \in \mathcal{D}$ and for any subset of output predictions (e.g., labels) $S \subseteq \mathcal{R}$, the following holds:

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\varepsilon} \Pr\left[\mathcal{M}\left(d'\right) \in S\right] + \delta. \qquad (1)$$

The definition above produces a notion of privacy that can be expressed as a measure of the probabilistic difference of a specific outcome by a multiplicative factor, e^{ε}, and an additive amount, δ. The DP mechanism with post-processing [16] (e.g., batch-wise training) is under a general Renyi-divergence [14] measurement with order $\alpha \in (1, \infty)$,

called RDP_α, for all neighboring data sets d, d':

$$\begin{aligned}&\text{RDP}_\alpha\left(\mathcal{M}(d)\|\mathcal{M}\left(d'\right)\right)\\&=\frac{1}{\alpha-1}\log S_{\theta\sim\mathcal{M}(d')}\left[\left(\frac{p_{\mathcal{M}(d)}(\theta)}{p_{\mathcal{M}(d')}(\theta)}\right)^{\alpha}\right]\leq\varepsilon\end{aligned}\tag{2}$$

As $\alpha\rightarrow\infty$, RDP_α converges to the standard $(\varepsilon,0)$-DP. Both ε and δ should be non-negative. Considering $\delta\rightarrow 0$ with only minor relaxation, a smaller value of ε indicates a stronger $(\varepsilon,0)$-differentially private guarantee. In other words, nearly equal probabilities in Eq. (1) would be given from the neighboring inputs d and d', which makes data identity much hard to infer. Moreover, learning from post-processing features (e.g., Mel-spectrum [30, 31]) based on the speech data could also be differentially private, which has been shown by the theorem given in [13, 16].

3. PATE LEARNING FOR ASR

3.1. Noisy Teacher Ensembles for Acoustic Modeling

We now describe PATE [18, 19] based acoustic modeling to enable ε-DP. First, an ensemble of teacher models is built by partitioning the training dataset into I disjoint subsets: $\mathcal{D}_1,...,\mathcal{D}_I$. Next, these are used to train I teacher models independently: $\mathcal{T}_1,...,\mathcal{T}_I$. Each such model is employed to generate frame-level acoustic model scores, which are then aggregated by weighted average and used as a teacher for student model training [32, 33, 34]. For each frame of audio x, the final aggregated teacher model produces a vector of posteriors $\mathcal{T}_{\text{ens}}(s\mid x)$ over context-dependent states s computed as follows:

$$\mathcal{T}_{\text{ens}}(s\mid x)=\sum_{i=1}^{I}w_i\mathcal{T}_i(s\mid x),\tag{3}$$

where $\mathcal{T}_i(s\mid x)$ is the posterior from the ith model, and w_i is its weighting coefficient. The corresponding states from individual teacher models have the same dimension J before the output alignment (e.g., considering a special silent character before alignment).

To ensure ε-DP under the PATE method, a random perturbation was introduced into the individual teachers' predictions ($\mathcal{T}_i$). We revise Eq. (3) to obtain a final ensemble from noisy teachers:

$$\mathcal{T}_{\text{PATE}}(x,\lambda)=\sum_{i=1}^{I}w_i\left(\mathcal{T}_i(x)+Y_j(\lambda)\right),\tag{4}$$

where $Y_1,...,Y_m$ are i.i.d. Laplacian or Gaussian random variables with location 0 and scale λ^{-1}. λ refers to a privacy parameter that influences (ϵ,δ)-DP guarantees and for which a bound has been proven under composition theorems applicable to model aggregation [18, 19, 33].

As shown in Figure 1, the next PATE step is a process of knowledge transfer, where the noisy ensemble output is used to relabel a nonsensitive data set, with a total sample number K, which in turn is used to train a student model, $\mathcal{S}$. Both the prediction outputs and the trained student model's internal parameters are free from querying requests, which allows the only privacy cost to be associated with acquiring the training data for the student model. We evaluate two noisy aggregation processes, Gaussian NoisyMax (GNMax) and Laplacian NoisyMax (LNMax) presented in previous studies [18, 19]. Under this setup, the student model is $(\varepsilon,10^{-3})$-DP guaranteed using λ from an analysis in [18, 19]. According to Eq. (4), a large λ refers to a **smaller ε, providing a stronger privacy guarantee**, but degrades the accuracy of the labels from the noisy maximum prediction output of the PATE function.

3.2. Sequence-level Teacher Distillations

To transfer the label data for knowledge distillation (KD) on a privacy-preserving student model, $\mathbf{h}^e_{\mathcal{T}}$ and $\mathbf{h}^e_{\mathcal{S}}$ denote the hidden vector representations of the teacher and student encoders respectively, and $P_{\mathcal{T}}\left(v\mid\mathbf{h}^d_{\mathcal{T}}\right)$ and $P_{\mathcal{S}}\left(v\mid\mathbf{h}^d_{\mathcal{S}}\right)$ the posterior probabilities computed by teacher and student models, respectively, in regards to the label v. $\mathcal{L}_{KD}$ is defined at the sequence-level as:

$$\mathcal{L}_{KD}=-\sum_{S}\sum_{n=1}^{N}P_{\mathcal{T}}\left(\hat{\mathbf{y}}_n\mid\mathbf{h}^e_{\mathcal{T}}\right)\log p_{\mathcal{S}}\left(\hat{\mathbf{y}}_n\mid\mathbf{h}^e_{\mathcal{S}}\right)\tag{5}$$

with $\hat{\mathbf{y}}_n$ being the nth hypothesis from the set of N-best hypothesis derived from beam-search (i.e., beam width $=N$) for the teacher.

3.3. Three Evaluated Acoustic Models

Three deep ASR models are adopted to investigate the impact of PATE. We consider homogeneously distributed settings, where the same network architecture is used in both teacher ($\mathcal{T}$) and student ($\mathcal{S}$) models in teacher-student learning.

RNN transducer (RNN-T) [20] utilizes a joint network to combine current acoustic observations from an encoder network and predictions based on previously recognized tokens. The encoder network is an RNN that converts the input acoustic features into a hidden representation for each frame of input. The prediction network generates output from previous nonblank output labels. The joint network computes output token logits. The blank symbol is treated as a possible output to account for the length mismatch between input and output token sequences, as in CTC [35] models. The loss function of RNN-T [20] is computed as the negative log posterior of the output label sequence given the input acoustic feature.

Hybrid CTC/attention (CTC/Att) [21] combines a sequence-to-sequence (seq2seq) attention-based [36] model with a CTC loss [35]. The encoder processes an input sequence and creates a hidden latent representation of the same length as the target sequence. The training data set contains target utterances composed of predefined vocabulary tokens (characters, tokens, or words). The CTC loss is computed from the predictions obtained from the encoder with sequence modeling.

LAS [22] consists of 3 components: (i) a **listener** (encoder), which is similar to a standard acoustic model, takes a time-frequency representation of the speech input and uses a group of neural network layers to map the input to higher-level features; (ii) an **attender**, which takes encoder features as input to learn an alignment between the input features and the predicted subword units, where each subword is typically a grapheme or word piece; and finally, (iii) **speller** (decoder), which takes the output of the attention module and generates a probability for each hypothesized word.

4. EXPERIMENTS

4.1. Experimental Setup

We consider two privacy-preserving conditions for ASR modeling with ε-DP measurement and protection:

1. Parts of the training data require ε-DP protection, and
2. All training data requires ε-DP protection,

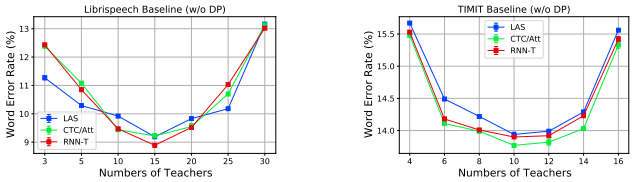
using two open-source speech corpora for training and evaluation: LibriSpeech [5] and TIMIT [37].

In Condition 1 ($\mathcal{C}1$), we assume both public and nonpublic data are from the same domain (e.g., same recording process and acoustic condition), but part of the training speech is associated with information that is potentially sensitive and needs protection. We select LibriSpeech and split the "train-clean960" set into 800 hours of "sensitive" data for learning teachers models, and 160 hours (as the nonsensitive data shown in Fig. 1) for training a ε-DP preserving student models with open access to end-users. We use the Librispeech "test-clean" portion to report word error rates (WERs).

In Condition 2 ($\mathcal{C}2$), we assume that two data sets coming from different domains can be accessed during acoustic modeling. One data set (LibriSpeech in our case) is publicly available in its entirety without privacy concerns, while the other data set (TIMIT in our case) is private and needs to be protected. We select different ASR models with pretraining on 800 hours of LibriSpeech used in $\mathcal{C}1$, and adopt the models for fine-tuning on the TIMIT training set (55 hours) as teacher models. The ε-DP preserving student ASR model is trained with a knowledge distillations process using the same unlabeled data as $\mathcal{C}1$. We use the 5-hour TIMIT validation data to report WERs.

4.2. Baseline and PATE Results

Ensemble baseline: Before incorporating the aggregated ASR system for noisy teacher-student learning, we first evaluate how the size of the teacher ensemble influences general ASR performance. In previous work on PATE, researchers have demonstrated that having sufficient training data for each teacher model is crucial to ensure a high-performance end-to-end model. We start with the same number of teacher models reported in [18, 19] with I=20 in Eq. (3) and evaluate with the various ASR models discussed in Section 3.3. The experimental results show that the data scale of the nonpublic training set used for teacher models could affect an optimal ensemble number. As shown in Fig. 2(a), the selected ASR models have lower WERs ($< 10\%$) with 10 to 20 teachers in $\mathcal{C}1$ (training with LibriSpeech 800 hours as private). For $\mathcal{C}2$ (training with TIMIT 55 hours as private) shown in Fig. 2(b), the desired number of teachers becomes smaller (10 to 12) to obtain competitive WER performance ($< 14\%$). We fix the number of teachers at 15 in $\mathcal{C}1$ and 10 in $\mathcal{C}2$, for all following PATE-based experiments.

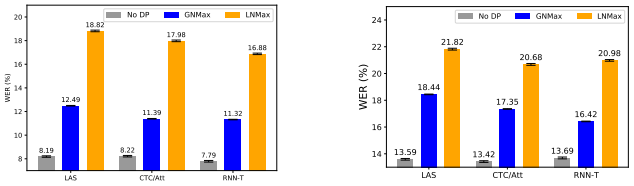


(a) $\mathcal{C}1$: LibriSpeech (b) $\mathcal{C}2$: TIMIT

Fig. 2: Baseline performance under ensemble training for ASR.

Additive noise selection: to study how the ε-DP based noise injection could impact the ASR performances, we compare two additive noise mechanisms, GNMax (with Gaussian noise) and LNMax (with Laplacian noise), with an empirical DP budget of ε=10, as has been reported for different on-device applications [38]. We follow the PATE framework proposed in Section 3 to obtain the final student ASR models with ε-DP protection. As shown in Fig. 3, both PATE-GNMax (blue) and PATE-LNMax (orange) show absolute WER degradation of 2.97% to 8.22% ($\mathcal{C}1$) and 2.73% to 8.23% ($\mathcal{C}2$) when compared to student model training with clean aggregated teacher models (gray). RNN-T based ASR with GNMax achieves the best performance under the ε=10 privacy budget, where PATE-GNMax shows an average of $4.32 \pm 0.12\%$ absolute WER reduction compared with PATE-LNMax results.

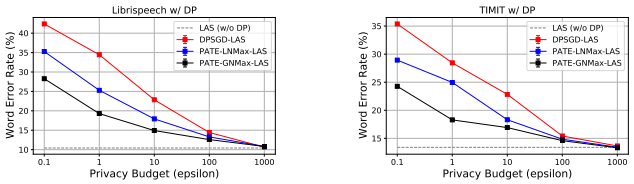
PATE for ASR with different privacy budgets: We evaluate the proposed PATE-based algorithms at different noise levels, which represent different target privacy budgets ($\lambda = \frac{K}{2\varepsilon}$). When compared with the benchmark ε-DP preserving DP-SGD [16] algorithm, the results in Figures 4 (LAS) 5 (Hybrid CTC/Att), and 6 (RNN-T) demonstrate that PATE under the GNMax mechanism can achieve good performance



(a) $\mathcal{C}1$: LibriSpeech (b) $\mathcal{C}2$: TIMIT

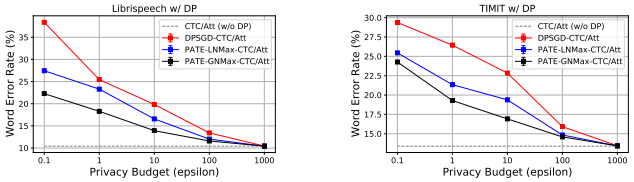
Fig. 3: Student ASR model performance with two additive noisy mechanisms, GNMax and LNMax, under PATE for teacher training.

when combined with different end-to-end architectures. They show $5.45 \pm 0.23\%$ and $9.81 \pm 0.43\%$ average WER reduction when comparing PATE-LNMax against DP-SGD [16] powered ε-DP preserving ASR models. They also highlight that the deployed models suffer major WER increases for privacy budgets $\varepsilon \leq 10$. The RNN-T based PATE model outperforms both LAS and Hybrid CTC/attention based ASR models when $\varepsilon \leq 100$. As an extreme case, the ASR model starts to converge to its clean ensemble when $\varepsilon \to 1000$ in both $\mathcal{C}1$ and $\mathcal{C}2$ conditions.



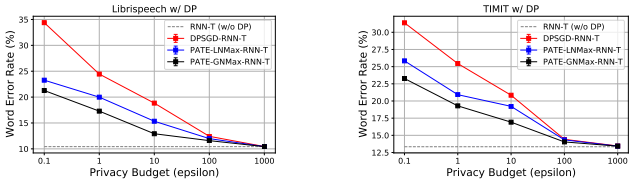
(a) $\mathcal{C}1$: LibriSpeech (b) $\mathcal{C}2$: TIMIT

Fig. 4: LAS results with PATEs and DP-SGD.



(a) $\mathcal{C}1$: LibriSpeech (b) $\mathcal{C}2$: TIMIT

Fig. 5: CTC/Attention results with PATEs and DP-SGD.

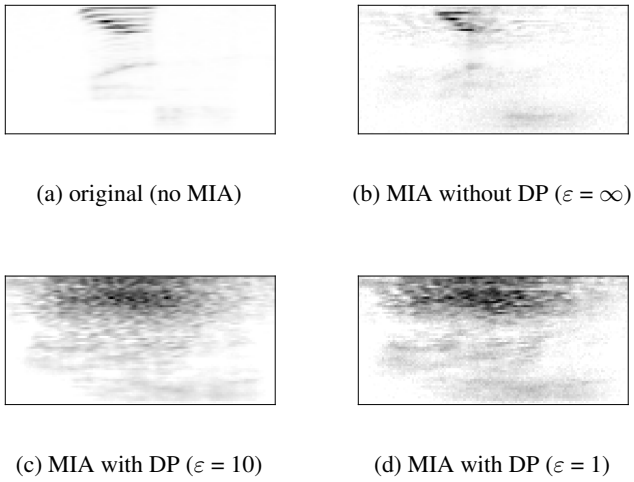


(a) $\mathcal{C}1$: LibriSpeech (b) $\mathcal{C}2$: TIMIT

Fig. 6: RNN-T results with PATEs and DP-SGD.

4.3. A Case Study for ε-DP Protection Against MIA

As a first step toward incorporating ε-DP into ASR acoustic modeling, we demonstrate how PATE can help protect user information in the training set, as used for training the teacher models. The demonstration considers the major privacy leakage method based on model inversion attack (MIA) [15], which utilizes the maximum a posteriori principle and constructs the input features that maximize the likelihood of observed model output queries (e.g., using an on-device API). We select the RNN-T based ensemble model trained under $\mathcal{C}1$ for a privacy-preserving case study against MIA with 10,000 service queries. We use the MIA algorithm to maximize the likelihood of a target output word of "stop" and observe its reconstructed high-confidence input audio. As shown in Fig. 7, MIA can generate inverse Mel-spectrum output as in Fig. 7(b), similar to an original utterance clip shown in Fig. 7(a), presumably including speech characteristics unique to the speaker. However, ε-DP shows its effectiveness in preventing this information recovery attack with $\varepsilon \leq 10$, as shown Fig. 7 (c) and Fig. 7 (d). This demonstrates how PATE-trained ASR models can successfully protect user information.



(a) original (no MIA) (b) MIA without DP ($\varepsilon = \infty$)

(c) MIA with DP ($\varepsilon = 10$) (d) MIA with DP ($\varepsilon = 1$)

Fig. 7: Model inversion attack (MIA) against ASR with a target word of "stop." (c) and (d) show effective ε-DP protection.

5. CONCLUSION

We have conducted a first study in privacy-preserving ASR applying the PATE framework for ε-differential privacy to three popular ASR architectures. For all the DP budgets tested, PATE greatly outperforms previously proposed DP-SGD mechanisms as benchmarks, especially under strict DP budgets (ε=1). Using the best RNN-T model, we also demonstrated that the proposed ε-DP preserving PATE models can protect against potential private data leakage from the training set using model inversion attacks.

6. REFERENCES

[1] Lawrence R. Rabiner and Ronald W. Schafer, *Introduction to Digital Speech Processing*, vol. 1, Now Publishers Inc, 2007.

[2] Ian McGraw, Rohit Prabhavalkar, Raziel Alvarez, Montse Gonzalez Arenas, Kanishka Rao, David Rybach, Ouais Alsharif, Haşim Sak, Alexander Gruenstein, Françoise Beaufays, et al., "Personalized speech recognition on mobile devices," in *Proc. IEEE ICASSP*, 2016, pp. 5955–5959.

[3] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[4] Janet M Baker, Li Deng, James Glass, Sanjeev Khudanpur, Chin-Hui Lee, Nelson Morgan, and Douglas O'Shaughnessy, "Developments and directions in speech recognition and understanding, Part 1 [DSP Education]," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 75–80, 2009.

[5] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*, 2015, pp. 5206–5210.

[6] Peng Cheng and Utz Roedig, "Personal voice assistant security and privacy—a survey," *Proceedings of the IEEE*, vol. 10, no. 4, pp. 476–507, 2022.

[7] Chao-Han Huck Yang, Sabato Marco Siniscalchi, and Chin-Hui Lee, "PATE-AAE: Incorporating adversarial autoencoder into private aggregation of teacher ensembles for spoken command classification," in *Proc. Interspeech*, 2021, pp. 881–885.

[8] Xiaodong Cui, Songtao Lu, and Brian Kingsbury, "Federated acoustic modeling for automatic speech recognition," in *Proc. IEEE ICASSP*, 2021, pp. 6748–6752.

[9] Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, et al., "The VoicePrivacy 2020 Challenge: Results and findings," *Computer Speech & Language*, vol. 74, pp. 101362, 2022.

[10] Chao-Han Huck Yang, Zeeshan Ahmed, Yile Gu, Joseph Szurley, Roger Ren, Linda Liu, Andreas Stolcke, and Ivan Bulyko, "Mitigating closed-model adversarial examples with Bayesian neural modeling for enhanced end-to-end speech recognition," in *Proc. IEEE ICASSP*, 2022, pp. 6302–6306.

[11] Haibin Wu, Xu Li, Andy T Liu, Zhiyong Wu, Helen Meng, and Hung-Yi Lee, "Improving the adversarial robustness for speaker verification by self-supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 202–217, 2021.

[12] Yu-Chen Lin, Tsun-An Hsieh, Kuo-Hsuan Hung, Cheng Yu, Harinath Garudadri, Yu Tsao, and Tei-Wei Kuo, "Speech recovery for real-world self-powered intermittent devices," in *Proc. IEEE ICASSP*, 2022, pp. 26–30.

[13] Cynthia Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*. Springer, 2008, pp. 1–19.

[14] Ilya Mironov, "Rényi differential privacy," in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. IEEE, 2017, pp. 263–275.

[15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1322–1333.

[16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[17] Arun Rajkumar and Shivani Agarwal, "A differentially private stochastic gradient descent algorithm for multiparty classification," in *Artificial Intelligence and Statistics*. PMLR, 2012, pp. 933–941.

[18] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *Proc. International Conference on Learning Representations*, 2017.

[19] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson, "Scalable private learning with PATE," in *Proc. International Conference on Learning Representations*, 2018.

[20] Alex Graves, "Sequence transduction with recurrent neural networks," *Representation Learning Workshop, ICML*, 2012.

[21] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE ICASSP*, 2017, pp. 4835–4839.

[22] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, Attend and Spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE ICASSP*, 2016, pp. 4960–4964.

[23] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan, "Boosting and differential privacy," in *Proc. IEEE 51st Annual Symposium on Foundations of Computer Science*, 2010, pp. 51–60.

[24] David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau, "Federated learning for keyword spotting," in *Proc. IEEE ICASSP*, 2019, pp. 6341–6345.

[25] Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Pin-Yu Chen, Sabato Marco Siniscalchi, Xiaoli Ma, and Chin-Hui Lee, "Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition," in *Proc. IEEE ICASSP*, 2021, pp. 6523–6527.

[26] Cornelius Glackin, Gerard Chollet, Nazim Dugan, Nigel Cannings, Julie Wall, Shahzaib Tahir, Indranil Ghosh Ray, and Muttukrishnan Rajarajan, "Privacy preserving encrypted phonetic search of speech data," in *Proc. IEEE ICASSP*, 2017, pp. 6414–6418.

[27] Dimitrios Dimitriadis, Kenichi Kumatani, Robert Gmyr, Yashesh Gaur, and Sefik Emre Eskimez, "A federated approach in training acoustic models," in *Proc. Interspeech*, 2020.

[28] Ferdinand Brasser, Tommaso Frassetto, Korbinian Riedhammer, Ahmad-Reza Sadeghi, Thomas Schneider, and Christian Weinert, "VoiceGuard: Secure and private speech processing.," in *Proc. Interspeech*, 2018, pp. 1303–1307.

[29] Manas A Pathak and Bhiksha Raj, "Privacy-preserving speaker verification and identification using Gaussian mixture models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 397–406, 2012.

[30] Steven Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[31] John S. Bridle and Michael D. Brown, "An experimental automatic word recognition system," *JSRU Report*, vol. 1003, no. 5, pp. 33, 1974, Joint Speech Research Unit Ruislip, England.

[32] Yevgen Chebotar and Austin Waters, "Distilling knowledge from ensembles of neural networks for speech recognition.," in *Proc. Interspeech*, 2016, pp. 3439–3443.

[33] Yan Gao, Titouan Parcollet, and Nicholas D Lane, "Distilling knowledge from ensembles of acoustic models for joint CTC-attention end-to-end speech recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, pp. 138–145.

[34] Chao-Han Huck Yang, Linda Liu, Ankur Gandhe, Yile Gu, Anirudh Raju, Denis Filimonov, and Ivan Bulyko, "Multi-task language modeling for improving speech recognition of rare words," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, pp. 1087–1093.

[35] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd International Conference on Machine Learning*, 2006, pp. 369–376.

[36] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE ICASSP*, 2016, pp. 4945–4949.

[37] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," vol. 93, pp. 27403, 1993.

[38] Wennan Zhu, Peter Kairouz, Brendan McMahan, Haicheng Sun, and Wei Li, "Federated heavy hitters discovery with differential privacy," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 3837–3847.