INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

PATE-AAE: Incorporating Adversarial Autoencoder into Private Aggregation of Teacher Ensembles for Spoken Command Classification

Chao-Han Huck Yang[1], *Sabato Marco Siniscalchi*[1,2,3], *Chin-Hui Lee*[1]

[1]Georgia Institute of Technology, Atlanta, GA, USA
[2]Faculty of Computer and Telecommunication Engineering, University of Enna, Italy
[3]Department of Electronic Systems, NTNU, Trondheim, Norway

{huckiyang,chl}@gatech.edu, marco.siniscalchi@unikore.it

Abstract

We propose using an adversarial autoencoder (AAE) to replace generative adversarial network (GAN) in private aggregation of teacher ensembles (PATE), a solution for ensuring differential privacy in speech applications. The AAE architecture allows us to obtain good synthetic speech leveraging upon a discriminative training of latent vectors. Such synthetic speech is used to build a privacy-preserving classifier when non-sensitive data is not sufficiently available in the public domain. This classifier follows the PATE scheme that uses an ensemble of noisy outputs to label the synthetic samples and guarantee ε-differential privacy (DP) on its derived classifiers. Our proposed framework thus consists of an AAE-based generator and a PATE-based classifier (PATE-AAE). Evaluated on the Google Speech Commands Dataset Version II, the proposed PATE-AAE improves the average classification accuracy by +2.11% and +6.60%, respectively, when compared with alternative privacy-preserving solutions, namely PATE-GAN and DP-GAN, while maintaining a strong level of privacy target at ε=0.01 with a fixed δ=10^{-5}.

Index Terms: Privacy-preserving speech processing, differential privacy, generative modeling, ensemble learning

1. Introduction

The speech signal contains a rich set of information [1] that encompasses gender, accent, speaking environment, and other speaker characteristics; therefore, protecting data privacy becomes a raising concern when speech data is used to deploy commercial speech applications. In recent years, public regulations, e.g., GDPR [2] and CCPA [3], have been proposed to establish new guidelines related to data privacy measurement and identity protection in end-user applications. Recent works on model inversion attacks [4, 5] indeed highlighted the importance of data privacy when the original data profile (e.g., facial images [4]) could be recovered from a machine learning model by using query-free optimization techniques.

Differential privacy [6] (DP) is an effective mechanism for ensuring individual data protection, and it has been deployed in several industrial systems [7, 8][1] to protect customer's sensitive information by exploiting a sophisticated noisy perturbation scheme. The ε-DP mechanism [6] provides a way to quantify a privacy loss and set up a privacy budget (e.g., a minimum ε value) for a given dataset. However, ε-differential private models [7] need to be refined in order to improve a degraded prediction accuracy [9] caused by the DP noise. The private aggregation of teacher ensembles [10] (PATE) is a recently proposed solution that aims to combat the accuracy loss of the machine learning models while ensuring privacy requirements. PATE follows a teacher-student architecture [11], where the teacher is an ensemble model. The underpinning idea in PATE is to leverage upon noisy outputs of aggregated teacher models to (re)label non-sensitive public data with DP guarantees. The PATE method and its improved version [12] were proven useful in reducing the model accuracy drop through a voting process during the noisy ensemble. Nonetheless, the teacher-student learning process highly depends on a hypothesis [10, 12, 13] that there exists a sufficient amount of public (non-sensitive) data to train the model. PATE-GAN [13] tries to overcome this issue by incorporating a generative block jointly trained with the PATE block; the goal is providing enough synthetic data to train deep models effectively. Unfortunately, PATE-GAN does not work well for high dimensional data synthesis (e.g., images), as demonstrated in recent studies [14, 15]. Moreover, generating speech samples is a challenging task, as shown in recent studies about neural vocoders [16, 17].

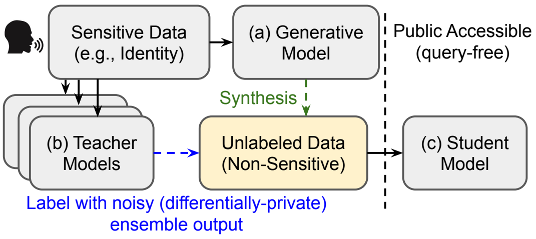


Figure 1: *Private aggregation of teachers ensemble (PATE) learning process [10, 12]: (a) the teacher prediction models training from sensitive data; (b) a joint generative model (e.g., adversarial autoencoder [18] for audio synthesis in this study); (c) student prediction model training from non-sensitive data.*

In this study, we introduce an adversarial autoencoder [18] (AAE) based model into PATE to improve the generative process for privacy-preserving speech classification. PATE-AAE first adapts an autoencoder to minimize a reconstruction loss, training on sensitive data. As shown in Fig. 1(a), the generative model produces synthetic data as non-sensitive samples. Meanwhile, the training data are divided into I isolated subsets to train individual teacher classifiers. For instance, I is equal to 3 in Fig. 1(b). The teacher classifiers then undergo an output aggregation process to generate noisy labels, which ensures the ε-differentially private protection. Finally, a student classifier uses the labeled synthetic samples (non-sensitive data) for training its model. The proposed PATE-AAE framework is

[1] Apple has also applied differential privacy with a privacy budget (ε=8) based on an official document in https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf.

http://dx.doi.org/10.21437/Interspeech.2021-640

assessed with the Google Speech Commands Dataset Version II [19]. Our experimental evidence demonstrates competitive results in terms of synthetic sample quality and classification accuracy with a strong ε-DP guarantee ($\varepsilon < 1$) considering established privacy-preserving learning (PPL) works [13, 20]. To the best of the authors' knowledge, this is the **first** attempt to introduce the PATE architecture into a speech classification task. Moreover, the proposed solution is benefited from adversarial autoencoder block, with advantages over existing GAN solutions [4, 13] of having a better test-likelihood estimation.

2. Related Work

Much of research effort to preserve data privacy in a machine learning model can be categorized into one of the two main groups: (i) systemic, such as federated learning [21], data isolation [22], and data encryption [23], and (ii) algorithmic, mainly differential private machine learning [7]. In the following sections, we first briefly discuss some of the privacy-preserving solutions proposed for speech applications. Next, we describe the substratum of differential privacy devised for machine learning applications and discuss the difference with our proposed approach while highlighting its key contributions.

2.1. Privacy-Preserving Speech Processing

Federated architectures [21, 24, 22] have been studied in the speech processing community to increase privacy protection. For example, the average gradient method [24] was used to update the learning model for decentralized training [25]. Heterogeneous computing architectures [22, 26] shows advantages on acoustic feature extraction for vertical federate learning. However, those approaches at a system-level usually make some assumptions on the limited accessibility of the malicious attackers and provide less universal measures about the privacy guarantees. There exist also some algorithmic efforts on investigating privacy-preserving speech processing by using cryptographic encryption [23, 27], and computation protocols [28]. Meanwhile, these encryption algorithms and protocols barely cover the training sample-level privacy protection, which plays a major role in deploying large-scale machine learning models.

However, differentially private algorithms [7, 29] is with a different focus from the aforementioned frameworks, aiming to provide quantitative guarantees and further prevent identity (e.g., accent) inference. We define a mathematical formation of differential privacy and investigate potential impacts on speech processing in the following sections.

2.2. Differential Privacy Fundamentals

The differential privacy mechanism [6] is an established standard 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., acoustic data) $d, d' \in \mathcal{D}$ and for any subset of outputs (e.g., labels) $S \subseteq \mathcal{R}$, the following holds:

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

The above definition provides a notion of privacy that can be interpreted as a measure of the probabilistic difference of a specific outcome by a multiplicative factor, e^{ε}, and an additive amount, δ. Both ε and δ should be positive or equal to zero. Considering $\delta \to 0$ 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 be inference. Moreover, learning from post-processing features (e.g., mel-frequency cepstral coefficients (MFCC)) from the data could also be differentially private, which have been proofed by the theorem given in [6].

2.3. Differential Privacy for Machine Learning

More recently, Abadi *et al.* [7] introduced the composition theorem [29], which guarantees the validity of DP protection when batch-wise training is used to learn deep neural network (DNN) parameters with non-convex objective functions. DP-GAN [20] incorporated noisy perturbations into a generative model during gradient updates to satisfy ε-DP but shows degraded prediction accuracy. Recent advances in teacher-student ensembles methods, such as PATE [10, 12], have further shown state-of-the-art performance on large-scale image classification tasks with sufficient non-sensitive data. Jordon *et al.* [13] instigated a GAN-based [30] generator into PATE to extend using scenarios with sufficient synthetic data (as non-sensitive), which is called PATE-GAN [13]. PATE-GAN is aligned with our motivation, but it only shows a stable performance for a small amount and low dimensional data. Meanwhile, application of PATE to large-scale speech processing, such as spoken-term classification, is practically absent despite the sensitive nature of the speech signals. In this study, we propose an autoencoder based approach incorporating PATE into speech processing, which considers non-sensitive acoustic data is not accessible. We will introduce an adversarial autoencoder with PATE to ensure ε-DP for the acoustic modeling in the next section.

3. PATE-AAE Framework

The proposed method consists of AAE and PATE models with feature encoders [31]. We focus on the application of PATE for speech processing, which is often in shortage of non-sensitive human voice data and more severe than in the original PATE [10]. It should be noted that PATE-GAN has succeeded in synthesizing low dimensional data (e.g., short sequences of EEG) but has failed when dealing with high-dimension data (e.g., images). This could be due to its difficulties of using a random noise generator to match *input data* distribution from *sample* discriminator in standard GAN [30] training.

3.1. Private Aggregation of Teacher Ensembles

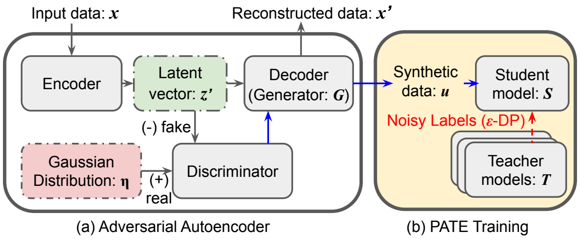
We describe the foundation [10, 12] of PATE to empower privacy-preserving speech classification. First, an ensemble of teacher models is built by partitioning the training dataset of images into n disjoint subsets: $\mathcal{D}_1, ..., \mathcal{D}_I$. Next, each subset is used to train I classifier independently: $\mathcal{T}_1, ..., \mathcal{T}_I$, that is, the teacher models. For each input data x, we aggregate prediction outputs from the teacher models to a single prediction. The number, $c_j(x)$, of teachers, that output class j for the given input x with m possible classes is set to be:

$$c_j(x) = |\{\mathcal{T}_i : \mathcal{T}_i(x) = j\}| \text{ for } j = 1, \ldots, m. \tag{2}$$

A random perturbation was introduced into the vote count, c_j, in Eq. (2) to obtain a noisy final prediction:

$$F_{\text{PATE}}(\mathbf{x}, \lambda) = \underset{j \in [m]}{\arg\max} \left(c_j(\mathbf{x}) + Y_j(\lambda)\right), \tag{3}$$

where $Y_1, ..., Y_m$ are i.i.d. $Lap(\lambda)$ random variables with location 0 and scale λ^{-1}. λ refers to a privacy parameter that influences (ϵ, δ)-differentially private guarantees and has been proven its bounded properties under composition theorems applying for model aggregation in [10, 12]. As shown in Figure 1, the next step in the PATE mechanism is based on a knowledge transfer process, where the noisy ensemble output is used to relabel a non-sensitive dataset, having a total sample number equal to K, which in turn is used to train a student model, $\mathcal{S}$. Both prediction outputs and the trained student model's internal parameters are free from querying requests, which allows the privacy cost only associated with acquiring the training data for the student model. Under the aforementioned data setup, the student model is $(\varepsilon, 10^{-5})$-differentially private guarantee using $\lambda = \frac{K}{2\varepsilon}$ from the analysis in [10, 12]. According to Eq. (3), a large λ refers to a smaller ε providing a strong privacy guarantee but degrade the accuracy of the labels from the noisy maximum prediction output of the PATE function.



(a) Adversarial Autoencoder (b) PATE Training

Figure 2: *The proposed PATE-AAE framework.*

3.2. PATE with Adversarial Autoencoder (PATE-AAE)

We introduce AAE training as follows. Instead of training the noisy generator as in GAN, AAE leverages upon an autoencoder-based regression model to minimize a reconstruction loss between input data (x) and decoded output (x'). The bottleneck vector (latent space z) is modeling as variational autoencoder [32] but uses a discriminator to refine z closing to a real vector (η) sampling from a fixed Gaussian distribution as shown in Fig. 2(a). This discriminative training resulted in a better test-likelihood on synthetic samples [18]. The recent success [31] of probabilistic autoencoder for audio synthesis motivates our solution that combines PATE with AAE for spoken command classification. Let x and z be the input and the bottleneck latent vector of an encoder-decoder model, respectively. The universal approximator posterior $q(z)$ introduced [18] is:

$$q(z \mid x) = \int_{\eta} q(z \mid x, \eta) p_{\eta}(\eta) d\eta \qquad (4)$$

$$\Rightarrow q(z) = \int_{x} \int_{\eta} q(z \mid x, \eta) p_{d}(x) p_{\eta}(\eta) d\eta dx, \qquad (5)$$

where the stochasticity in $q(z)$ comes from both the data-distribution x and the random noise η with a fixed Gaussian distribution at the input of the encoder. The adversarial training procedure can match $q(z)$ to $p(z)$ by back-propagation through the encoder network directly. The encoder encodes an input data x into latent vector, $z'_i \sim \mathcal{N}(\mu_i(\mathbf{x}), \sigma_i(\mathbf{x}))$, by variational inference used in [32]. Therefore, a training objective for reconstructing input x is computed by minimizing the following upper-bound on the negative log-likelihood of x:

$$E_x \left[\mathrm{E}_{q(z|\mathbf{x})}[-\log(p(x \mid z)] \right] + E_x[\mathrm{KL}(q(z \mid x) \| p(z))]. \quad (6)$$

Makhzani *et al.* [18] further introduce a discriminative update into the second terms of Eq. (6) that makes $q(z)$ to match to the distribution of $p(z)$ to train an AAE. The discriminative training objective between the latent vector z' (denoted as a fake sample as 0) and the sampling noise η (denoted as a real sample as 1) is computed by BCE loss and back-propagated gradients to input data (x) for updating encoder's parameters (Fig. 2 (a)).

Next, for training teacher models, we partition the sensitive dataset into n subsets, $\mathcal{D}_1, \ldots, \mathcal{D}_I$, with $|\mathcal{D}_i| = \frac{|\mathcal{D}|}{I}$ for $\forall q$. Each teacher model (T_i) is training with discriminator loss:

$$\mathcal{L}_{T_i} = -(\sum \log T_i(q(z)) + \sum_{j=1}^{I} \log(1 - T_i(q(z'_j)))) \quad (7)$$

To generate synthetic samples for training student model, we take I samples of Gaussian distribution, $\eta_1, ..., \eta_I$ using the trained AAE decoder network (G) to synthesize sample $\hat{\mathbf{u}}_j = G(\eta_j)$ for each class. Following the PATE mechanism for knowledge transfer, the aggregated noisy output from the teachers models in Eq. (7) labels the synthetic data for training a differentially private student model, where noisy label refers $r_j = \mathrm{PATE}(\hat{\mathbf{u}}_j, \lambda)$ from Eq. (3). Finally, we train the student model to maximize the standard cross-entropy loss on this teacher-labeled data:

$$\mathcal{L}_S = \sum_{j=1}^{I} r_j \log S(\hat{\mathbf{u}}_j) + (1 - r_j) \log(1 - S(\hat{\mathbf{u}}_j)) \qquad (8)$$

We train $G, T_1, ..., T_I$ and S iteratively, with each iteration of G consisting of first performing gradient updates on all teachers, then performing gradient updates of the student. The major difference between proposed PATE-AAE and PATE-GAN [13] is on the generative process. Proposed AAE method uses a regression autoencoder architecture to reconstruct the input samples and use a random variable for the refined latent space as the decoder (generator) input for generating new synthetic samples. Instead, PATE-GAN adapts discriminator on the sample generator directly to refine the learning process from random noise to synthetic data, which produces worse test-likelihood on high dimensional data from previous studies [14, 33] related to the convergence properties [34] of GAN. To conduct privacy-preserving speech processing frameworks, we select PATE-GAN [13], and DP-GAN [20] as baselines in our studies motivated by the condition of without any available non-sensitive audio dataset.

4. Experiment

4.1. Experimental Setup

(1) Dataset and Classifier Model: A large-scale dataset (~100k training samples) for the partition process is needed in our experiments. Therefore, we have chosen the Google Speech Commands V2 [19] task, which contains 105,829 utterances of 35 words from 2,618 speakers with a sampling rate of 16 kHz. The audio length per sample clip is 1 second for a total amount of 55.5 hours. We split the dataset into I=200 [13] disjoint subsets to train individual teacher classifiers following the procedure indicated in Eq. (7). We use the mel-spectrogram feature with an 80-band mel-scale and 1024 points of discrete Fourier transform as inputs to the classifiers. For a fair comparison, both teacher and student classifiers use an identical self-attention [35] and U-Net [36] based neural network proposed in the recent work [22], which has shown benchmark prediction accuracy on the selected speech commands dataset.
(2) Encoder-Decoder Model: For encoder-decoder inputs, we

use standard 13 MFCC at a sampling rate of 100 Hz from 80 log-mel filter-bank features on training (x) and synthetic (u) data. We carefully build our encoder-decoder upon a WaveNet-based autoencoder presented in [31], which has shown competitive performance for unsupervised speech synthesis tasks. As shown in Fig 3(a), our encoder has four network blocks associated 768 input units, which include the first ResNet layer, a convolution layer with a stride scale = 2, the second ResNet layer, and a Dense layer with ReLU activation. The latent representation z is computed by outputs (μ, σ) from two linear layers with 128 units. As shown in Fig 3(b), our decoder applied a randomized dropout layer from an output of the discriminator, which is selected from a latent vector (z) or a random vector (η). The output vector is then upsampled 320 times to fit the 16 kHz sampling rate for WaveNet decoding. Finally, we follow the same setting with [31] for running two cycles of WaveNet with 20 convolution layers followed by a 256-ReLU layer. We follow μ-law companding transformation [37, 17] with 256 quantization levels to generate raw 16k Hz audio.

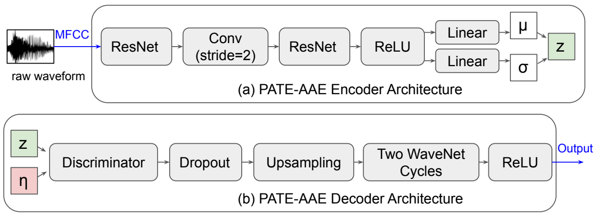
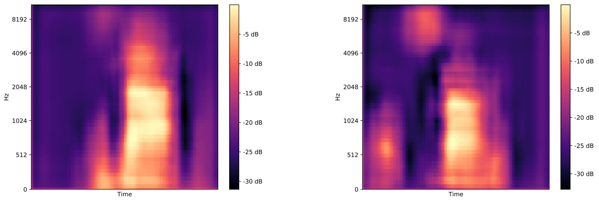


Figure 3: *PATE-AAE encoder-decoder architecture.*

(3) Synthesis and Classification Evaluation Metrics: For the speech synthesis task, the Frechet Inception Distance (FID) [38] is selected to evaluate the sample quality that computes the Frechet Distance [39] between two multivariate Gaussian distributions for the synthetic and real samples. We follow a standard FID setup in [15] to evaluate the quality of over 10,000 synthetic speech samples generated from random noise. For the speech command classification task, classification accuracy is used to evaluate the student model. We train each privacy-preserving model 20 times and report its average prediction accuracy.

Table 1: *FID scores (lower is better) for speech synthesis quality by generative models with ε-DP (a fixed δ=10^{-5}) settings.*

privacy target (ε)	0.01	0.1	1	10	100
DP-GAN [20]	35.2±0.5	32.2±0.6	28.8±0.4	26.7±0.3	24.9±0.3
PATE-GAN [13]	33.4±0.5	30.1±0.4	27.9±0.3	26.0±0.3	24.3±0.1
PATE-AAE	**30.2**±0.4	**28.6**±0.3	**26.3**±0.2	**24.7**±0.2	**24.0**±0.1



(a) *PATE-GAN*

(b) *PATE-AAE*

Figure 4: *Mel-spectrogram of (a) PATE-GAN; (b) PATE-AAE with a privacy target (ε=0.1), where output command is "right."*

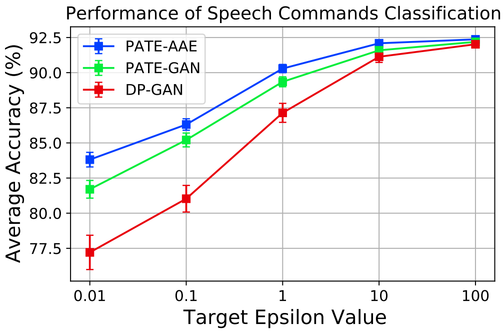


Figure 5: *Performance of different privacy-preserving models under four levels of privacy target (ε, with a fixed δ=10^{-5}), which refers to the Laplace noise level at scale of $\lambda = \frac{K}{2\varepsilon}$.*

4.2. Results and Performance Analysis

We have developed two baseline systems, namely PATE-GAN [13] and DP-GAN [20], whose models are using the same encoder-decoder architecture introduced in Sec. 4.1.(2). First, as shown in Tab. 1, PATE-AAE performs best (lowest) FID scores compared with PATE-GAN and DP-GAN, which indicates a good capability of privacy-preserving speech synthesis. According to a visualization presented in Fig. 4 (a), PATE-AAE demonstrates many detailed acoustic features on its mel-spectrogram, where PATE-GAN's mel-spectrogram at most preserves intensity information. Next, we have investigated the prediction accuracy in terms of different target privacy budgets with the constraint of guaranteeing $(\varepsilon, 10^{-5})$-differential privacy. The classification results are averaged over 20 trials to reduce the effect of random parameters initialization.

Fig. 5 summarises our results. As a fair comparison, a visual inspection of Fig. 5 reveals that the two baseline systems and the proposed PATE-AAE system attain a similar average classification accuracy with an extremely weak privacy budget (ε=100). As the constraints on the privacy increases, that is, ϵ is reduced (the noisy level, $\lambda = \frac{K}{2\varepsilon}$, is increased). Simultaneously, the proposed PATE-AAE approach exemplifies its advantage over both the DP-GAN and PATE-GAN solutions. In particular, PATE-AAE boosts the average accuracy by 2.11% (ε=0.01) and 1.11% (ε=0.1) compared with PATE-GAN; 6.60% (ε=0.01) and 5.32% (ε=0.1) compared with DP-GAN. With a large ε value, the noise level (λ) becomes too small; nonetheless, PATE-AAE still attains a slightly better average accuracy (92.37%) than PATE-GAN (92.19%) and DP-GAN (92.02%) baselines. This phenomenon could be due to the aggregation in PATEs ameliorating the negative impact of ε-DP noise and echoing theoretical studies in [12, 13, 14].

5. Conclusion

In this paper, we incorporate an adversarial autoencoder into the PATE scheme. We further investigate different privacy-preserving solutions to speech command classification by combining the WaveNet based encoder-decoder structures and classification models together. The proposed PATE-AAE approach shows the best performances in terms of synthetic speech quality scores and classification accuracy. Our future work includes exploring different ε-DP distributed training strategies, such as average gradient aggregation [14] and adversarial training [40], for large vocabulary continuous speech recognition.

6. References

[1] M. A. Pathak, B. Raj, S. D. Rane, and P. Smaragdis, "Privacy-preserving speech processing: cryptographic and string-matching frameworks show promise," *IEEE signal processing magazine*, vol. 30, no. 2, pp. 62–74, 2013.

[2] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, p. 3152676, 2017.

[3] S. Shatz and S. E. Chylik, "The california consumer privacy act of 2018: A sea change in the protection of california consumers'," *The Business Lawyer*, vol. 75, 2020.

[4] M. Fredrikson, S. Jha, and T. 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.

[5] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models," *arXiv preprint arXiv:2012.07805*, 2020.

[6] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[7] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[8] D. Kifer, S. Messing, A. Roth, A. Thakurta, and D. Zhang, "Guidelines for implementing and auditing differentially private systems," *arXiv preprint arXiv:2002.04049*, 2020.

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

[10] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *ICLR*, 2017.

[11] H. Hu, S. M. Siniscalchi, Y. Wang, and C.-H. Lee, "Relational teacher student learning with neural label embedding for device adaptation in acoustic scene classification," *Proc. Interspeech 2020*, pp. 1196–1200, 2020.

[12] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and U. Erlingsson, "Scalable private learning with pate," in *International Conference on Learning Representations*, 2018.

[13] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in *International Conference on Learning Representations*, 2019.

[14] D. Chen, T. Orekondy, and M. Fritz, "Gs-wgan: A gradient-sanitized approach for learning differentially private generators," *Neural Information Processing Systems (NeurIPS)*, 2020.

[15] K. N. Haque, R. Rana, and B. W. Schuller, "High-fidelity audio generation and representation learning with guided adversarial autoencoder," *IEEE Access*, vol. 8, pp. 223 509–223 528, 2020.

[16] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[17] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[18] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[19] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[20] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," *arXiv preprint arXiv:1802.06739*, 2018.

[21] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6341–6345.

[22] C.-H. H. Yang, J. Qi, S. Y.-C. Chen, P.-Y. Chen, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition," *arXiv preprint arXiv:2010.13309*, 2020.

[23] C. Glackin, G. Chollet, N. Dugan, N. Cannings, J. Wall, S. Tahir, I. G. Ray, and M. Rajarajan, "Privacy preserving encrypted phonetic search of speech data," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 6414–6418.

[24] D. Dimitriadis, K. Kumatani, R. Gmyr, Y. Gaur, and S. E. Eskimez, "A federated approach in training acoustic models," in *Proc. Interspeech*, 2020.

[25] J. Qi, C.-H. H. Yang, and J. Tejedor, "Submodular rank aggregation on score-based permutations for distributed automatic speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3517–3521.

[26] S. Y.-C. Chen and S. Yoo, "Federated quantum machine learning," *arXiv preprint arXiv:2103.12010*, 2021.

[27] F. Brasser, T. Frassetto, K. Riedhammer, A.-R. Sadeghi, T. Schneider, and C. Weinert, "Voiceguard: Secure and private speech processing." in *Interspeech*, vol. 18, 2018, pp. 1303–1307.

[28] M. A. Pathak and B. 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.

[29] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010, pp. 51–60.

[30] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[31] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[32] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[33] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "Gan-leaks: A taxonomy of membership inference attacks against generative models," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 343–362.

[34] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, "On convergence and stability of gans," *arXiv preprint arXiv:1705.07215*, 2017.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[37] C. Recommendation, "Pulse code modulation (pcm) of voice frequencies," in *ITU*, 1988.

[38] K. Shmelkov, C. Schmid, and K. Alahari, "How good is my gan?" in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 213–229.

[39] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of multivariate analysis*, vol. 12, no. 3, pp. 450–455, 1982.

[40] C.-H. Yang, J. Qi, P.-Y. Chen, X. Ma, and C.-H. Lee, "Characterizing speech adversarial examples using self-attention u-net enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3107–3111.