SPEECH PRIVACY LEAKAGE FROM SHARED GRADIENTS IN DISTRIBUTED LEARNING

Zhuohang Li[1], *Jiaxin Zhang*[2], *Jian Liu*[1]

[1]University of Tennessee, Knoxville [2]Intuit AI Research

ABSTRACT

Distributed machine learning paradigms, such as federated learning, have been recently adopted in many privacy-critical applications for speech analysis. However, such frameworks are vulnerable to privacy leakage attacks from shared gradients. Despite extensive efforts in the image domain, the exploration of speech privacy leakage from gradients is quite limited. In this paper, we explore methods for recovering private speech/speaker information from the shared gradients in distributed learning settings. We conduct experiments on a keyword spotting model with two different types of speech features to quantify the amount of leaked information by measuring the similarity between the original and recovered speech signals. We further demonstrate the feasibility of inferring various levels of side-channel information, including speech content and speaker identity, under the distributed learning framework without accessing the user's data.

***Index Terms*—** distributed learning, privacy leakage, speech processing

1. INTRODUCTION

Voice assistants, such as Google Assistant, Amazon Alexa, and Apple Siri, have been widely deployed on various smartphones and smart speakers, as they provide a natural and convenient way for user interaction. The modern voice user interface is powered by deep neural networks, which enables efficient speech processing for many tasks, such as *automatic speaker verification* (ASV) [1] and *automatic speech recognition* (ASR) [2]. The remarkable performance of such models is fueled by a growing amount of training data; yet collecting data from users is becoming increasingly difficult due to privacy regulations [3, 4] and user privacy concerns.

Distributed machine learning, which allows multiple data holders to jointly train a machine learning model under the coordination of a central server, has attracted great research attention. Compared with the conventional centralized learning framework, data parallelism distributed training not only scales better to larger data sizes, but also provides a level of privacy to participating users by encoding the data minimization principle: clients are allowed to keep their private data on the local device and only share the gradient information to the server for updating the model. As such, distributed learning, especially the emerging *federated learning* (FL), has been quickly deployed in production for many speech-related tasks, such as speaker verification [5] and keyword spotting [6].

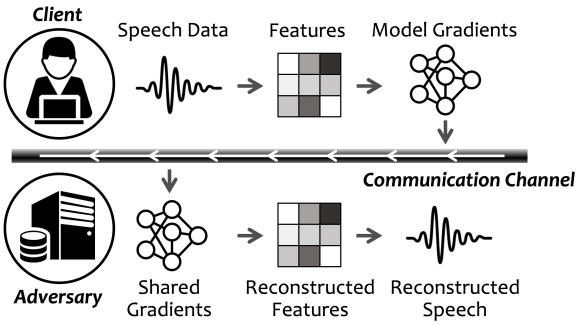


Fig. 1. Illustration of speech privacy leakage from shared gradients in the distributed learning scenario.

Recently, several studies [7, 8, 9, 10] have revealed that image data can be recovered to some extent through the shared gradients in distributed learning (known as *gradient leakage*, or *gradient inversion*), which may pose severe threats to user's data privacy. However, to date, there has been limited exploration of gradient leakage in the speech domain. Compared with image data, speech recordings are a rich source of personal and sensitive information that can be used to support a wide spectrum of applications, from sentiment analysis to spoken language recognition, and further to voice biometric profiling. Therefore, distributed learning involving speech data should be evaluated carefully to fully understand its potential privacy vulnerabilities.

To fill this gap, this work investigates the risk of gradient leakage on speech data with the following questions:

1. How to recover private speech data from the shared gradients, if feasible?
2. What level of private information (e.g., speech content, speaker identity, etc.) can be exposed from gradients?

To answer the first question, we extend on previous gradient leakage attacks in the image domain and provide a two-stage inversion method that can numerically restore the speech waveform from the gradients shared by the client, which is illustrated in Fig. 1. We find that different from the image domain, deep learning models for processing speech data are usually trained on condensed speech features rather than the raw waveform. As a result, inverting the gradients

only recovers the features of the original signal. Moreover, unlike the image data which has well-defined value space for each pixel, the spectral/cepstral feature of speech signals has a wider range and is more prone to subtle errors which would be amplified when projecting back to the time domain, and thus would require more careful treatment.

To answer the second question, we design and conduct extensive experiments to investigate the amount of information that can be recovered through the gradients. We explore a distributed learning scenario of speech command recognition on two types of features, i.e., Mel-spectrogram and Mel-frequency cepstral coefficients (MFCC), and utilize a suite of 4 different metrics to quantify the recovered speech quality and the intelligibility of the recovered speech content. Moreover, we further inspect the amount of leaked voice biometric information leveraging a pre-trained speaker verification model. Our results show that compared with MFCC, using Mel-spectrogram as the front-end feature would lead to more leakage in speech content and speaker information of the client's private speech data from the shared gradients. These findings can help the community understand the potential of gradient leakage on different speech features under the gradient sharing framework and design distributed learning schemes with enhanced privacy protection.

2. RELATED WORK

Distributed Learning for Speech. Recently, an increasing amount of research effort has been put into utilizing federated learning as a privacy-enhancing technique to improve neural networks on distributed user devices, such as smartphones and smart speakers. Besides early studies [11, 12, 13] conducted in simulated environments, many federated learning frameworks for speech processing have already been deployed in production, including Apple's speaker verification model [5] and Google's keyword spotting model [6]. As human speech contains rich semantics of sensitive and personal information, there is a pressing need for thoroughly understanding the privacy risks of models trained on speech data in a distributed manner.

Privacy Leakage from Gradients. A few recent studies have shown that sharing gradients in distributed learning is not as safe as it has been presumed. Zhu *et al.* [7] first demonstrated that it is possible to recover clients' private images from the shared gradients by generating fake images that minimize the gradient matching loss. Geiping *et al.* [8] extended this method to deeper neural networks with higher resolution images with a different loss function design. Following work [9, 10] further improved on this by exploring various prior information. Compared to the remarkable progress in the image domain, there is a lack of research on the feasibility and severity of gradient leakage in speech data. To our knowledge, the most relevant study by Dang *et al*. [14] proposed a method to infer speaker identity from gradients of an ASR model. However, such a method only recovers the speech features instead of the original speech waveform. Moreover, compared with ASR models that are trained on long spoken sentences, the lightweight keyword spotting models for recognizing speech commands are more commonly deployed in distributed learning setting, yet the privacy risk there is still unexplored. This work is devoted to filling this research gap.

3. METHODOLOGY

3.1. Problem Formulation

We consider a supervised learning task under the canonical distributed learning setting that involves two parties: the server S and the clients C. The learning objective is to optimize the parameters θ of a neural network f_θ to minimize the empirical risk measured by loss function $\mathcal{L}$ on all training data: $\min_\theta \sum_{c\in C} \sum_{(x_i,y_i)\in\mathcal{D}_c} \mathcal{L}(f_\theta(x_i), y_i)$, where x_i, y_i are the local data and label from client c's local dataset $\mathcal{D}_c$. Instead of directly sharing their private data, federated learning allows participating clients to only shared the gradients computed on their private data, i.e., $\nabla_\theta \mathcal{L}(f_\theta(x_i), y_i)$. The server then collects and aggregates gradients from all participating clients to update the global model for each communication round.

Threat Model. We assume the adversary cannot interfere with the normal federated learning procedure but has access to the gradients uploaded by each individual client. In practice, the adversary can be an honest-but-curious server or a malicious analyst that eavesdrops on the communication channel.

Objective. The adversary's objective is to infer private information about the client by attempting to reconstruct the client's private data. Previous studies [7, 8] have shown that this can be done through generating synthetic data samples to match the client's gradients. Let $\Delta\theta$ denote the actual gradients shared by the client and x', y' represent the synthetic data sample and label. Formally the adversary solves for:

$$x'^*, y'^* = \underset{x',y'}{\operatorname{argmin}} \mathbf{d}(\Delta\theta, \nabla_\theta \mathcal{L}(f_\theta(x'), y')), \qquad (1)$$

where x'^*, y'^* is the minimizer of the distance between gradients measured by a distance metric $\mathbf{d}$. Intuitively, if the gradients computed on the synthetic data closely match the actually shared gradients, the synthetic data will also recover the important semantics of the client's private data. In practice, as the label y'^* can be analytically restored from the gradients [8, 9], the adversary only needs to solve for x'^*.

3.2. Recovering Speech Data From Gradients

Challenge. In the image domain, deep learning models are usually trained to directly process raw imagery data. Differently, in the speech domain, it is a common practice to first extract spectral or temporal acoustic features (e.g., spectrogram or cepstral coefficients) from the raw speech signal and then feed the extracted features into the deep learning model.

This creates an additional layer of difficulty to gradient inversion since solving Eq. 1 only recovers the features rather than the original speech waveform.

Method. To address the above challenge, the proposed method recovers speech data from the shared gradients in the following two stages:

(1) *Feature Reconstruction*. The goal of the first stage is to recover the acoustic features from the gradients shared by the users. Specifically, let u denote the set of 2-dimensional spectral features extracted from speech waveform x. We then solve the optimization problem in Eq. 1 by minimizing the Euclidean distance in the model parameter (gradient) space:

$$u'^* = \underset{u'}{\operatorname{argmin}} ||\Delta\theta - \nabla_\theta \mathcal{L}(f_\theta(u'), y'^*)||_2^2 + \lambda \mathbf{r}(u'), \quad (2)$$

where $\mathbf{r}$ is a regularization term and $\lambda > 0$ is a weighting parameter. In this work, we use anisotropic total variation [15] as the regularizer, i.e., $\mathbf{r}(u') = ||d_h u'||_1 + ||d_v u'||_1$, where d_h, d_v denote the horizontal and vertical partial derivative operators, respectively. We do not bound the search space for h' during optimization since unlike image data, the values of the acoustic features do not reside in a well-defined range.

(2) *Waveform Reconstruction*. The goal of the second stage is to reconstruct the speech waveform based on the features recovered from the first stage. In this work, we consider recovering the waveform from two common types of features for speech processing: Mel-spectrogram and MFCC. To convert a Mel-spectrogram back to a time-domain signal, we first create Mel filter banks and then approximate the normal short-time Fourier transform (STFT) magnitudes by searching for the non-negative least squares solution that minimizes the Euclidian distance between the target Mel-spectrogram and the product of the estimated spectrogram and the filter banks. Then the Griffin-Lim algorithm [16] is applied to produce the speech waveform by estimating the missing phase information. As for MFCCs, an extra step is needed to first invert the cepstral coefficients to approximate a Mel-spectrogram. This is done by first applying the inverse discrete cosine transform (iDCT) and then map the decibel-scaled results to a power spectrogram[1]. After that, the regular Mel-spectrogram inversion procedure is applied to further get the estimated waveform.

4. EXPERIMENTS

4.1. Experimental Setting

Dataset. We use speech data from the Speech Commands dataset [17] to conduct evaluation. The dataset was developed for developing and testing compact and effcient on-device keyword spotting model, which is one of the most widely-adopted federated learning applications in production, and thus is well-suited for our task. Each sample of the dataset contains 1 second recording of spoken speech commands sampled at 16kHz and corresponding label. We select a subset of the dataset that contains 10 common words (i.e., "yes", "no", "up", "down", "left", "right", "on", "off", "stop", and "go") as in the first released version of the dataset.

Table 1. Model used in evaluation.

Type	Kernel	Stride	Output
Conv2D	(3, 3)	(1, 1)	(30, 30, 32)
Conv2D	(3, 3)	(1, 1)	(28, 28, 64)
MaxPooling2D	(2, 2)	(2, 2)	(14, 14, 64)
Flatten	–	–	12544
FC	–	–	128
FC	–	–	10

Front-end Feature Extraction. To extract acoustic features, the speech signal is first pre-emphasised with a factor of 0.97. Then speech frames are created using overlapping Hamming windows with length of $2,048$ samples and frame-shift of 512 samples. For *Mel-spectrogram*, 512 point fast Fourier transform (FFT) with 32 Mel bands is used. For *MFCC*, 128-channel filterbank is used to extract 32-dimensional coefficient features which are further processed by cepstral mean and variance normalization (CMVN).

Distributed Learning Setting. (1) *Model*. We adopt the same model structure as used in the Tensorflow keyword recognition example[2]. The model is composed of two 2D convolution layers, one max-pooling layer, and two fully-connected layers. The detailed model architecture is described in Table 1. (2) *Gradient Computation*. We assume the clients compute one local step of gradient descent on one random sample from their private dataset and then send back the gradients (i.e., distributed stochastic gradient descent).

Parameters. We solve Eq. 2 using Adam optimizer for $8,000$ iterations with $\lambda = 0.001$ and a learning rate of 0.01. To avoid local minimum, each sample is given 2 trials and the reconstruction result with the lower loss is selected.

4.2. Experimental Results

4.2.1. Reconstruction Evaluation

To evaluate the reconstruction performance, we compare the resulting speech signal recovered from gradients using our method with the original signal using the following metrics:

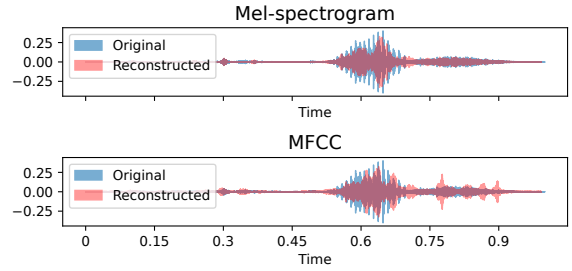
Evaluation Metrics. (1) *F-MSE*: the mean squared error between the ground truth features of the original speech and the reconstructed features. (2) *W-MSE*: the mean squared error between the original speech waveform and the reconstructed waveform. (3) *PESQ*: the perceptual evaluation of speech quality (PESQ) [18] score is designed for end-to-end quality assessment of degraded audio sample in narrow-band telephone networks. The computed score is in the range of $[-0.5, 4.5]$, with higher scores indicating better speech

[1] https://librosa.org/doc/main/generated/librosa.feature.inverse.mfcc_to_mel.html

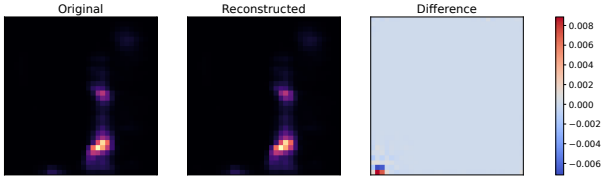
[2] https://www.tensorflow.org/tutorials/audio/simple_audio

Table 2. Reconstruction results on 400 speech samples.

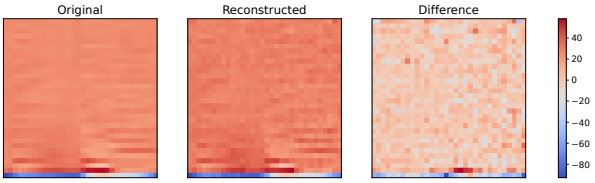
	F-MSE ↓	W-MSE ↓	PESQ ↑	STOI ↑
Inverting from ***Features***				
Mel-spectrogram	–	0.0097 ± 0.0130	2.1090 ± 0.3563	0.8082 ± 0.0715
MFCC	–	0.0091 ± 0.0119	2.1041 ± 0.3762	0.7856 ± 0.0760
Inverting from ***Gradients***				
Mel-spectrogram	0.0002 ± 0.0017	0.0095 ± 0.0129	2.0427 ± 0.3816	0.8004 ± 0.0804
MFCC	5701.6225 ± 1919.7741	0.0153 ± 0.0086	1.3886 ± 0.1750	0.4259 ± 0.1234



(a) Waveform



(b) Mel-spectrogram



(c) MFCC

Fig. 2. Visualizing reconstructed speech command "yes".

quality. (4) *STOI*: the short-time objective intelligibility (STOI) [19] metric measures the intelligibility of degraded speech signals based on a correlation coefficient between the temporal envelopes of the reference and degraded signals in short-time overlapping segments.

Results. Table 2 presents the statistical results of conducting the reconstruction on 400 random speech samples from the testing set. We also show the results of inverting from the ground truth features (i.e., only conducting waveform reconstruction) as baseline for comparison. An example of visualizing the reconstructed speech command "yes" is provided in Fig. 2. We observe that for Mel-spectrogram, inverting from gradients yields a similar performance as directly inverting from features, and the reconstructed waveform is very close to the original waveform, with measured W-MSE < 0.001 and PESQ > 2. However, for MFCC, inverting from gradients would induce a large distortion, causing the quality of the reconstructed speech to degrade drastically. This is potentially because the coefficient values are in decibel scale and has high variance and thus are prone to small perturbations, which makes it harder to launch gradient leakage attacks against MFCCs.

Table 3. Speaker re-identification results on 400 speech samples.

	Score ↑	Success Rate ↑
w.r.t. ***Inverted Signal from Features***		
Mel-spectrogram	0.7288 ± 0.1386	99.25%
MFCC	0.1574 ± 0.1119	16%
w.r.t. ***Original Signal***		
Mel-spectrogram	0.4445 ± 0.1450	90.5%
MFCC	0.0514 ± 0.0916	2.5%

4.2.2. Speaker Re-identification

To examine whether the speaker information (i.e., voice biometric) can be retained through the reconstruction, we pass the signal recovered from gradients and the reference signal into a speaker verification model and report the cosine similarity score of the embeddings and the success rate of the two signals being recognized as the same speaker. To perform speaker verification, we use the ECAPA-TDNN model [20] pretrained on Voxceleb dataset provided by SpeechBrian[3].

Results. Table 3 presents the statistical results of conducting speaker re-identification on the same set of 400 speech samples. For comparison, we show the results measured w.r.t. both the inverted signal from the ground truth features and w.r.t. the original signal. We observe that the speech signals reconstructed from Mel-spectrogram preserve most speaker information, with 99% and 90% chance to pass the speaker verification when compared to the inverted signal and original signal, respectively. In contrast, speech signals reconstructed from MFCC have a very low probability to be verified as the same speaker, especially when directly compared to the original signal.

5. CONCLUSION

In this work, we study the potential of the privacy leakage of speech data from shared gradients by proposing a two-stage inversion method that sequentially achieves speech feature reconstruction from gradients and waveform reconstruction from the recovered features. Through extensive experiments, we demonstrate that compared to Mel-spectrogram, MFCC exhibits better resilience against gradient leakage attacks, with less leaked speech/speaker information. Future work could investigate neural vocoders for better waveform reconstruction quality.

Acknowledgement. This work was supported in part by NSF CNS-2114161, ECCS-2132106, CBET-2130643, and the Science Alliance's StART program.

[3] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

6. REFERENCES

[1] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification.," in *Interspeech*, 2017, vol. 2017, pp. 999–1003.

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[3] European Union, "General Data Protection Regulation," `https://gdpr-info.eu/`, 2021.

[4] Office of the California Attorney General, "California Consumer Privacy Act (CCPA): Proposed Text of Regulations," `https://www.oag.ca.gov/sites/all/files/agweb/pdfs/privacy/ccpa-proposed-regs.pdf`, 2021.

[5] Filip Granqvist, Matt Seigel, Rogier van Dalen, Aine Cahill, Stephen Shum, and Matthias Paulik, "Improving on-device speaker verification using federated learning with privacy," 2020.

[6] Andrew Hard, Kurt Partridge, Neng Chen, Sean Augenstein, Aishanee Shah, Hyun Jin Park, Alex Park, Sara Ng, Jessica Nguyen, Ignacio Lopez Moreno, et al., "Production federated keyword spotting via distillation, filtering, and joint federated-centralized training," *arXiv preprint arXiv:2204.06322*, 2022.

[7] Ligeng Zhu, Zhijian Liu, and Song Han, "Deep leakage from gradients," *Advances in neural information processing systems*, vol. 32, 2019.

[8] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller, "Inverting gradients-how easy is it to break privacy in federated learning?," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[9] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov, "See through gradients: Image batch recovery via gradinversion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16337–16346.

[10] Zhuohang Li, Jiaxin Zhang, Luyang Liu, and Jian Liu, "Auditing privacy defenses in federated learning via generative gradient leakage," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10132–10142.

[11] David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau, "Federated learning for keyword spotting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6341–6345.

[12] Andrew Hard, Kurt Partridge, Cameron Nguyen, Niranjan Subrahmanya, Aishanee Shah, Pai Zhu, Ignacio Lopez Moreno, and Rajiv Mathews, "Training keyword spotting models on non-iid data with federated learning," 2020.

[13] Dhruv Guliani, Françoise Beaufays, and Giovanni Motta, "Training speech recognition models with federated learning: A quality/cost framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3080–3084.

[14] Trung Dang, Om Thakkar, Swaroop Ramaswamy, Rajiv Mathews, Peter Chin, and Françoise Beaufays, "A method to reveal speaker identity in distributed asr training, and how to counter it," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4338–4342.

[15] Yifei Lou, Tieyong Zeng, Stanley Osher, and Jack Xin, "A weighted difference of anisotropic and isotropic total variation model for image processing," *SIAM Journal on Imaging Sciences*, vol. 8, no. 3, pp. 1798–1823, 2015.

[16] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

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

[18] John G Beerends, Andries P Hekstra, Antony W Rix, and Michael P Hollier, "Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part ii: psychoacoustic model," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, 2002.

[19] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[20] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, Helen Meng, Bo Xu, and Thomas Fang Zheng, Eds. 2020, pp. 3830–3834, ISCA.