Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

How to Recover Long Audio Sequences Through Gradient Inversion Attack With Dynamic Segment-based Reconstruction

Xijie Zeng[1,2], *Frank Rudzicz*[1,2]

[1]Dalhousie University, Canada
[2]Vector Institute, Canada

Xijie.Zeng@dal.ca, frank@dal.ca

Abstract

Recent advancements in gradient inversion attacks have demonstrated the vulnerability of shared gradients in distributed learning systems, particularly in image and text domains. However, applying these techniques to audio data, especially longer speech sequences, remains largely unexplored. Our study introduces a novel approach that builds upon the principles of gradient inversion attacks to retrieve high-quality audio recordings from shared gradients. We propose an optimized spectrogram segmentation technique that enables extracting longer audio sequences with diverse acoustic features, without requiring complex post-processing techniques. Through this study, we overcome the limitations of previous methods that were restricted to short audio clips with simple acoustic features and limited semantic information.

Index Terms: Data privacy, speech recognition, federated learning, Adversarial attack

1. Introduction

Voice interactions with smart home devices, virtual assistants, and voice messaging systems are becoming more common, and compromising user voice or speech information could significantly impede regular activities and violate personal privacy [1–3]. When audio data is used for sensitive applications such as voice authentication or health monitoring, the potential impact of privacy breaches becomes particularly concerning. In federated learning scenarios, protecting client training data from the server side and other participants in the distributed training process is paramount [4]. However, the foundational work on gradient inversion attacks, specifically Deep Leakage from Gradients (DLG) [5] and its improved version, iDLG [6], demonstrated that reconstructing training images from shared gradients in federated learning settings was possible. Recent studies [7–10] have also shown the capability of gradient inversion attacks to retrieve highly confidential training data based on shared gradients in computer vision and natural language processing tasks under federated learning settings.

Building upon these foundations, several studies incorporated regularization strategies to improve reconstruction quality. GradInversion [7], e.g., used batch normalization statistics as a strong image prior along with group consistency regularization and achieved high-resolution (224 $\times$ 224) image reconstruction. CAFE [11] incorporates regularization based on internal feature representations to raise the maximum batch size up to 100. To further improve the reconstruction quality on high-resolution images (from 64 $\times$ 64 to 256 $\times$ 256), subsequent works [8, 9, 12] explored using generative models in gradient inversion attacks. These approaches employed GANs [13] to improve the naturalness of reconstructed images. While they showed promise in producing more realistic reconstructions on higher resolution images, especially for complex datasets, many of these sophisticated methods required significant computational resources compared to the foundational DLG and iDLG approaches [5, 6], which could run efficiently on CPUs.

As gradient inversion attacks matured in the image domain, researchers began to explore their applicability to other modalities, including audio [14, 15]. They adapted optimization frameworks and regularization strategies [5, 16] from image reconstruction tasks, with the variation that audio attacks involve recovering spectral features and converting them back to waveforms. Those works demonstrated that reconstructed audio could reveal sensitive information like speech content and speaker identity, and that reconstructing from mel-spectrograms yielded better results than from MFCCs. However, they also focused only on datasets with short audio recordings like Speech Commands [17], AudioMNIST [18], and Urbansound [19], typically around 1 second in length, as their methods were not designed to handle the increased complexity of longer audio recordings. While these methods proved valuable as proofs-of-concept, they are inadequate for real-world scenarios where conversational speech recordings are often significantly longer, and more privacy-sensitive if reconstructed accurately. Recovering longer audio sequences also poses unique challenges that differ from short audio clips. Directly applying prior gradient inversion methods to long audio sequences — similar to reconstructing high-resolution images from aggregated gradients — fails to converge due to the exponentially growing parameter space of large mel-spectrograms. These factors make the reconstruction task substantially more demanding than simply scaling up or repeating the approach used for short clips. Hence, there is a significant gap in the current state of audio gradient inversion attacks for longer audio sequences.

We build upon and extend this line of research, focusing on high-fidelity reconstruction of long audio sequences from shared gradients. Our key contributions are:

- We develop a novel approach inspired by gradient inversion techniques that recovers audio sequences across a wide range of temporal lengths (from 1 second to over 2 minutes) in both individual and batch settings, while maintaining high fidelity. To the best of our knowledge, our work demonstrates the first successful high-quality reconstruction of clips up to 140 seconds in length.
- We achieve state-of-the-art reconstruction quality, with an average cosine similarity of 0.98 between original and reconstructed audio, significantly surpassing previous methods.
- We demonstrate the scalability of our approach across audio samples of varying lengths, sampling frequencies, noise categories, number of speakers, and identity information.

10.21437/Interspeech.2025-244

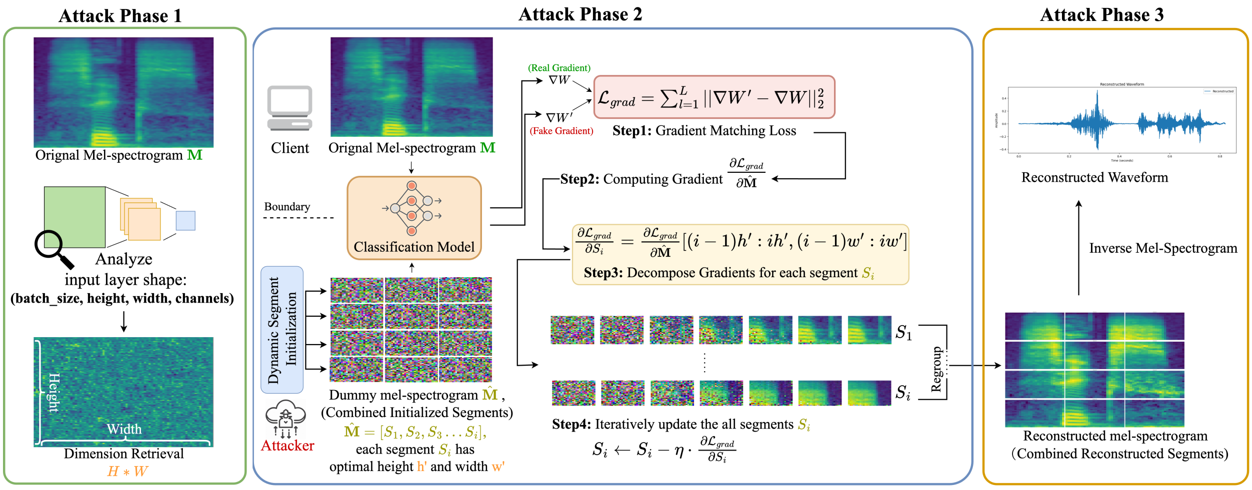

Figure 1: ***Attack Phase 1:*** *Reconstruction begins with an initial gradient analysis to determine the width and height of the input mel-spectrogram through the network's weight matrices.* ***Attack Phase 2:*** *Based on these dimensions, we initialize segments of optimal size for individual reconstruction through gradient matching optimization.* ***Attack Phase 3:*** *The reconstructed segments are regrouped to form the full mel-spectrogram, which is then converted back to a waveform.*

2. Methodology

2.1. Attack Overview

The reconstruction pipeline begins by analyzing shared gradients to extract the full mel-spectrogram's dimensions, which can be inferred from the network's first layer weight matrices available in the shared gradient information. Note that we assume the white-box attack scenario as in previous works [5–7, 14, 15], where the attacker has access only to the network architecture described in Section 3.2 and the shared gradients. To overcome the limitations of existing gradient inversion attacks in extracting long audio sequences with high-resolution mel-spectrograms, we propose a **dynamic segment initialization strategy** that determines and initializes segments with appropriate dimensions to ensure convergence of the gradient matching optimization. This process is implemented through two main functions: *Calculate_Segment_Size*, which determines optimal segment dimensions, and *Initialization*, which initializes segments based on these parameters. The reconstructed segments are then reassembled and converted back to waveform via inverse mel-spectrogram transformation. Figure 1 illustrates the complete reconstruction process.

Calculate_Segment_Size takes two parameters: the shape of the input mel-spectrogram recovered from the initial analysis, and a target segment size. It first establishes the segment height by taking the minimum between the target height and the actual spectrogram height. For the segment width, it iteratively searches for the largest divisor of the spectrogram width that falls within a predefined range. This ensures that the resulting segments maintain a suitable aspect ratio for reconstruction while enabling complete coverage of the input spectrogram without information loss.

The *Initialization* function systematically initializes segments based on the calculated dimensions while preserving their spatial ordering. Using the dimensions determined by *Calculate_segment_Size*, the function generates appropriately sized segments that will collectively form the complete mel-spectrogram when reconstructed. Each segment is initialized independently but maintains its relative position in the overall structure, ensuring proper alignment during the final reassembly phase.

2.2. Audio Feature Reconstruction

Single Audio Feature Recovery. After determining the segment dimensions, we apply our reconstruction approach to recover all initialized mel-spectrogram segments in parallel. We uses a classification model with cross-entropy loss to generate gradients that indicate how the model's parameters should change to improve its predictions.

The key insight of our method lies in analyzing the gradients of the final fully connected layer. Inspired by Zhao *et al.* [6] and Yin *et al.* [7], we observe that gradient patterns in this layer contain structural information that can guide reconstruction, which is an optimization process that iteratively updates the randomly initialized input segment $\hat{\mathbf{M}}$ to produce gradients most similar to those observed during the attack:

$$\hat{\mathbf{M}} = \arg\min_{\hat{\mathbf{M}}} \sum_{l=1}^{L} ||\nabla \mathbf{W}'_l - \nabla \mathbf{W}_l||_2^2, \tag{1}$$

where $\nabla \mathbf{W}'_l$ represents the gradients computed using the current estimate at layer l, and L is the total number of layers. We use L-BFGS optimizer [20] to iteratively update the mel-spectrogram segments that would generate gradients matching those observed during the attack. Figure 2 illustrates this retrieval process.

Multiple Audio Features Recovery. We also propose a new approach that extends to recover multiple audio recordings simultaneously in a batch. Given a batch of N mel-spectrograms $M_1, M_2, ..., M_N$, each mel-spectrogram M_i is divided into segments $M_i \rightarrow C_{i1}, C_{i2}, \dots, C_{ik_i}$, where k_i is the number of segments for the I^{th} audio file, each segment C_{ij} having dimensions (h_i, w_i). The process begins with normalizing each segment:

$$C_{ij}^{norm} = \frac{C_{ij} - \min(C_{ij})}{\max(C_{ij}) - \min(C_{ij})}. \tag{2}$$

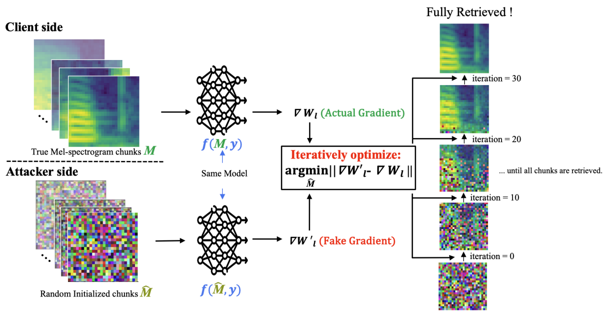

Figure 2: *Segments retrieval process for a single audio. The top illustrates the client-side training process where true mel-spectrograms generate actual gradients through forward and backward propagation. The bottom shows the attacker's reconstruction process, where an optimally initialized fake mel-spectrogram is iteratively updated by minimizing the Euclidean distance between fake gradients and actual gradients.*

We then form the batch tensor B and the dummy tensor $\hat{B}$:

$$B = [C_{11}^{norm}, ..., C_{1k_1}^{norm}, ..., C_{Nk_N}^{norm}], \hat{B} \sim \mathcal{U}(0, 1). \quad (3)$$

The objective is to find $\hat{B}$ that produces similar gradients:

$$\begin{aligned} \hat{B}^* = \arg\min_{\hat{B}} \sum_{l=1}^{L} ||\nabla \mathcal{L}(f(\hat{B}), y) - \nabla \mathcal{L}(f(B), y)||_2^2 \\ + \lambda_1 \mathcal{L}spec(\hat{B}, B) + \lambda_2 \mathcal{L}smooth(\hat{B}), \end{aligned} \quad (4)$$

where $\mathcal{L}$ is the cross-entropy loss and y is a dummy label. $\mathcal{L}spec$ is the spectral loss:

$$\mathcal{L}spec(\hat{B}, B) = |\mathcal{F}(\hat{B}) - \mathcal{F}(B)|_2^2, \quad (5)$$

where $\mathcal{F}$ is the Fourier transform. The smoothness loss is:

$$\mathcal{L}smooth(\hat{B}) = |\nabla_x \hat{B}|_1 + |\nabla_y \hat{B}|_1. \quad (6)$$

From Mel-spectrogram to audio signals. The final stage converts the reconstructed mel-spectrogram back into a waveform. We first convert the mel-spectrogram from decibel scale to power scale then compute the inverse mel-spectrogram with librosa[1], which employs the Griffin-Lim algorithm for phase reconstruction. Finally, we apply a low-pass Butterworth filter [2] to the resulting audio to reduce high-frequency artifacts.

3. Experiment Setup

3.1. Datasets

To evaluate our approach's effectiveness across diverse audio characteristics, we conducted experiments on several datasets. For single audio retrieval, we used: 400 audio clips chosen from the Speech Commands V2 test set [17], comprising 10 spoken commands ('yes', 'no', 'up', 'down', 'left', 'right', 'on', 'off', 'stop', and 'go') - following the selection in [14]) with 40 clips per command from 20 speakers; 100 samples from AudioMNIST [18], containing 10 samples for each spoken digit (0-9) from each of 10 speakers; 100 clips from UrbanSound8K [19] with 10 samples from each type of environmental sound; 50 random clips from the CommonVoice [21] and 50 clips from the LibriSpeech [22], using 5 clips from each of 10 speakers.

3.2. Global Model

Single Audio Retrieval. Following the same hypothesis as Li *et al.* [14] and Ovi *et al.* [15], we adopt a CNN architecture consists of four sequential convolutional layers, each with 12 output channels and 5×5 kernel size. The first two convolutional layers use a stride of 2, while the latter two maintain a stride of 1. Each followed by sigmoid activation.

Multiple Audio Retrieval. An enhanced LeNet-style [23] architecture is implemented to process mel-spectrogram segments from multiple audio recordings, following an encoder-decoder structure with residual connections. The encoder includes four convolutional layers (32, 64, 128, and 128 channels) with 3×3 kernels and padding of 1. Each convolutional layer is followed by LeakyReLU ($\alpha = 0.2$) and batch normalization, with two residual connections after the first and second convolutional layers. The decoder mirrors this structure with four transposed convolutional layers using the same kernel size and padding, followed by a Sigmoid activation.

3.3. Feature Preparation

We assume clients transform raw audio into mel-spectrograms following standard practices in audio processing, maintaining the native sampling rate. The preprocessing involves generating mel-spectrograms with 128 mel bins using a 2048-point Fast Fourier Transform (FFT) and a hop length of 512 samples, following common configurations such as those in the librosa library[3]. The resulting mel-spectrograms are converted to decibel scale before analysis.

4. Results

4.1. Retrieved Audio Quality Evaluation

We evaluate the quality of recovered audio clips using three metrics: Signal-to-Noise Ratio (SNR), Cosine Similarity between original and recovered mel-spectrograms (M-Cos.Sim.), and mean squared error between original and recovered waveforms (W-MSE). Tables 1 and 2 present the evaluation results of our proposed method. We compare single audio recovery with previous approaches across all datasets and evaluate multiple audio recovery using the same metrics, with batch sizes of 4, 8, and 16 for longer audio clips from CommonVoice and LibriSpeech, and up to 32 for shorter audio clips from SpeechCommands, AudioMNIST, and UrbanSound8k.

4.2. WER & Speaker Verification

To assess both semantic content preservation and speaker identification accuracy, we employ two specialized evaluation methods. First, we use Whisper (Large) [24] to generate transcripts of both original and recovered audio clips, computing the Word Error Rate (WER) between them. Second, we use SpeechBrain's ECAPA-TDNN [25] to compute two metrics: the mean cosine similarity score (Sim.Score) between original and recovered audio clips, and the mean probability that both audio samples originate from the same speaker (S.S.Prob.). Note that we conducted these evaluations using audio clips retrieved through the maximum batch sizes with lower audio quality. Table 3 presents these results.

[1]https://librosa.org/doc/main/generated/librosa.feature.inverse.mel_to_audio
[2]https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.butter
[3]https://librosa.org/doc/main/generated/librosa.feature.melspectrogram

Table 1: *Single Audio Recovery Results Across 5 Datasets. S = number of samples. $\hat{L}$ = average length of audio clips (seconds)*

Single Recovery			
Matrics	**SNR ↑**	**M-Cos.Sim. ↑**	**W-MSE ↓**
SpeechCommands (S=400, $\hat{L}$=0.98s)			
Li et al. [14]	32.451±3.254	0.632±0.126	0.027±0.032
Ours	**47.731±7.572**	**0.991±0.007**	**0.010±0.012**
AudioMNIST (S=100, $\hat{L}$=0.65s)			
Ovi et al. [15]	15.129±3.258	0.287±0.023	0.158±0.069
Ours	**35.814±4.325**	**0.982±0.005**	**0.003±0.001**
UrbanSound8k (S=100, $\hat{L}$=3.60s)			
Ovi et al. [15]	12.236±4.083	0.132±0.076	0.434±0.136
Ours	**21.107±5.461**	**0.961±0.018**	**0.021±0.025**
CommonVoice (S=50, $\hat{L}$=6.16s)			
Ours	**46.412±7.072**	**0.989±0.002**	**0.005±0.002**
LibriSpeech (S=50, $\hat{L}$=8.58s)			
Ours	**38.098±2.078**	**0.985±0.004**	**0.003±0.002**
Longest audio Recovery			
CommonVoice (139.43s audio)			
Ours	**40.098**	**0.992**	**0.023**
LibriSpeech (34.96s audio)			
Ours	**37.598**	**0.989**	**0.005**

Table 2: *Batch Audio Recovery Metrics Across 5 Datasets (SpeechCommands-SC, AudioMNIST-AMN, UrbanSound8k-US8k, CommonVoice-CV, Librispeech-Libr). B = batch size.*

Batch Recovery			
Datasets	**SNR ↑**	**M-Cos.Sim. ↑**	**W-MSE ↓**
SC(B=8)	23.772±4.003	0.988±0.002	0.013±0.005
SC(B=16)	33.415±5.105	0.982±0.007	0.013±0.007
SC(B=32)	**36.205±10.021**	0.995±0.002	0.019±0.013
AMN(B=8)	**26.270±1.186**	0.989±0.002	0.015±0.004
AMN(B=16)	21.463±1.565	0.988±0.003	0.012±0.005
AMN(B=32)	20.429±4.409	0.989±0.002	0.012±0.005
US8k(B=8)	17.601±2.973	0.975±0.010	0.040±0.030
US8k(B=16)	**18.694±4.534**	0.981±0.008	0.038±0.031
US8k(B=32)	17.839±2.188	0.972±0.016	0.024±0.025
CV(B=4)	**37.390±2.437**	0.994±0.001	0.007±0.003
CV(B=8)	33.031±2.244	0.994±0.001	0.007±0.002
CV(B=16)	32.192±2.263	0.993±0.002	0.007±0.002
Libr(B=4)	**24.752±2.035**	0.978±0.003	0.005±0.003
Libr(B=8)	21.570±1.618	0.970±0.007	0.005±0.003
Libr(B=16)	20.840±1.418	0.980±0.007	0.006±0.002

Table 3: *WER and Speaker Verification Results When Audios were Retrieved in Batch. B = batch size.*

Datasets	**WER ↓**	**Sim.Score ↑**	**S.S.Prob. ↑**
AMN (B=32)	-	**0.664±0.058**	**0.790±0.019**
SC (B=32)	-	0.359±0.127	0.670±0.059
CV (B=16)	**0.189±0.248**	0.522±0.081	0.739±0.032
Libr (B=16)	0.409±0.390	0.230±0.102	0.612±0.048

4.3. Results Analysis

Our approach demonstrates substantial improvements in single audio recovery compared to existing methods. For the SpeechCommands dataset, we achieve significantly higher SNR (47.731±7.572 dB compared to 32.451±3.254 dB) and mel-spectrogram cosine similarity (0.991±0.007 compared to 0.632±0.126) than previous work. Similar improvements appear across AudioMNIST and UrbanSound8k datasets. For longer-duration datasets (CommonVoice and LibriSpeech), our method maintains robust performance with SNR values of 46.412±7.072 dB and 38.098±2.078 dB respectively, and mel-spectrogram cosine similarities consistently above 0.98. The consistently low W-MSE values (below 0.01) across all datasets indicate high-quality waveform reconstruction.

Notably, our method successfully reconstructs considerably longer audio sequences while maintaining high quality. For a 139.43-second CommonVoice audio clip, we achieve an SNR of 40.098 dB and mel-spectrogram cosine similarity of 0.992. Similarly, for a 34.96-second LibriSpeech audio clip, we obtain an SNR of 37.598 dB and an impressive mel-spectrogram cosine similarity of 0.989. These results demonstrate our method's capability to handle audio sequences that are significantly longer than those attempted in previous studies.

In batch processing scenarios, short audio clips (SpeechCommands, AudioMNIST) maintain high reconstruction quality even with larger batch sizes, though SNR values show some variation. Longer clips (CommonVoice, LibriSpeech) perform optimally with smaller batch sizes, showing gradual quality reduction as batch size increases. Both single and batch recovery performed on UrbanSound8k show lower quality metrics due to the dataset's diverse acoustic characteristics, including varying background noise levels and overlapping urban sound sources.

The trade-off between batch size and reconstruction quality is reflected in the speaker verification and WER results. We notice that shorter utterances maintain high speaker similarity scores (0.664±0.058 for AudioMNIST), while longer recordings show moderate degradation (similarity score 0.230±0.102 for LibriSpeech). WER results on longer recordings (0.189±0.248 for CommonVoice, 0.409±0.390 for LibriSpeech) indicate reasonable preservation of semantic content despite the challenging conditions.

5. Conclusion

We have presented a novel approach for high-quality reconstruction of both individual and batched audio sequences from shared gradients, introducing a dynamic segment initialization strategy combined with principles from gradient inversion attacks. Our method significantly outperforms previous approaches that were limited to short clips, revealing critical privacy vulnerabilities in federated learning systems as longer audio clips typically contain more sensitive information such as speaker identity and extended speech content. Future work should explore vulnerabilities in other features (e.g., MFCC) and develop defense mechanisms for long audio protection, perhaps based on gradient perturbation or differential privacy.

6. Acknowledgements

Rudzicz is supported by a Canada CIFAR Chair in AI and a Killam Memorial Chair.

7. References

[1] J. Lau, B. Zimmerman, and F. Schaub, "Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, Nov. 2018. [Online]. Available: https://doi.org/10.1145/3274371

[2] P. Cheng and U. Roedig, "Personal voice assistant security and privacy—a survey," *Proceedings of the IEEE*, vol. 110, no. 4, pp. 476–507, 2022.

[3] J. Li, C. Chen, M. Rahimi Azghadi, H. Ghodosi, L. Pan, and J. Zhang, "Security and privacy problems in voice assistant applications: A survey," *Computers Security*, vol. 134, p. 103448, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404823003589

[4] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," 2016. [Online]. Available: https://arxiv.org/abs/1602.05629

[5] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/60a6c4002cc7b29142def8871531281a-Paper.pdf

[6] B. Zhao, K. R. Mopuri, and H. Bilen, "iDLG: Improved deep leakage from gradients," 2020. [Online]. Available: https://arxiv.org/abs/2001.02610

[7] H. Yin, A. Mallya, A. Vahdat, J. M. Alvarez, J. Kautz, and P. Molchanov, "See through gradients: Image batch recovery via gradinversion," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 332–16 341.

[8] J. Jeon, J. Kim, K. Lee, S. Oh, and J. Ok, "Gradient inversion with generative image prior," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Curran Associates Inc., 2024.

[9] A. Hatamizadeh, H. Yin, H. R. Roth, W. Li, J. Kautz, D. Xu, and P. Molchanov, "GradViT: Gradient inversion of vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 021–10 030.

[10] L. H. Fowl, J. Geiping, S. Reich, Y. Wen, W. Czaja, M. Goldblum, and T. Goldstein, "Decepticons: Corrupted transformers breach privacy in federated learning for language models," in *The Eleventh International Conference on Learning Representations*, 2023.

[11] X. Jin, P.-Y. Chen, C.-Y. Hsu, C.-M. Yu, and T. Chen, "CAFE: Catastrophic data leakage in vertical federated learning," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 994–1006.

[12] Z. Li, J. Zhang, L. Liu, and J. Liu, "Auditing privacy defenses in federated learning via generative gradient leakage," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 122–10 132.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[14] Z. Li, J. Zhang, and J. Liu, "Speech privacy leakage from shared gradients in distributed learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[15] P. R. Ovi and A. Gangopadhyay, "Gradient inversion attacks on acoustic signals: Revealing security risks in audio recognition systems," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 4835–4839.

[16] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients - how easy is it to break privacy in federated learning?" in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Curran Associates Inc., 2020.

[17] P. Warden, "Speech Commands: A dataset for limited-vocabulary speech recognition," 2018. [Online]. Available: https://arxiv.org/abs/1804.03209

[18] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapuschkin, and W. Samek, "AudioMNIST: Exploring explainable artificial intelligence for audio analysis on a simple benchmark," 2023. [Online]. Available: https://arxiv.org/abs/1807.03418

[19] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[20] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical Programming*, vol. 45, pp. 503–528, 1989. [Online]. Available: https://api.semanticscholar.org/CorpusID:5681609

[21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518.

[25] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech 2020*, 2020, pp. 3830–3834.