

Frames in the span: exact recovery of long-form speech from federated ASR updates

Taka Khoo

Independent researcher

Manuscript in preparation for ESANN 2027. Not yet peer reviewed.

Abstract. Federated speech recognition shares model updates in place of audio. We show that for models whose first layer applies one linear map to overlapping windows of feature frames, a single update contains the features. The rows of the first-layer gradient span the windows, the overlap makes membership in that span a linear system, and a banded solve returns the utterance: 98.5% of 1,417 LibriSpeech utterances up to 6.2s at 115 dB SNR, without transcript, length, loss or optimisation. The system is solvable up to $(k-s)F$ windows. Deeper layers extend the attack to 35s. It survives local SGD, batching, dropout and training.

1 Introduction

Federated learning (FL) trains a model on client devices and returns updates to a server [1]. For automatic speech recognition (ASR) the appeal is plain: the audio never leaves the device [2, 3]. Gradient inversion [4, 5] asks how much of the input the update still reveals.

For speech the answers so far come from optimisation. Dang et al. [6] reconstruct DeepSpeech [7] features with a zeroth-order search, given the transcript and length of short utterances, and identify the speaker in 34% of cases; dropout brings this to 0%. Later work matches gradients of small classifiers [8, 9], reaching long audio by segmenting it [10], or matches DeepSpeech gradients through the CTC loss [11, 12]. These attacks are slow, need side information, and their failures under dropout, local training or long inputs are read as evidence that realistic FL is hard to invert [13].

We take the algebraic route opened for text and images by DAGER [14], SPEAR [15] and TIGER [16]: the gradient of a linear layer is low rank and its row space is the span of the layer’s inputs. Our observation is that speech front ends make this span unusually easy to read. Frames of one utterance act as a batch, and each frame is seen through a *window* that overlaps its neighbours. The overlap is what turns span membership into a linear system. Contributions:

- a closed-form attack on the first layer with a capacity law, $(k-s)F$ windows for kernel k , stride s and F features, verified on four front ends;
- a sequential decoder that uses the spans leaked by deeper layers and recovers every test utterance length we tried, up to 35.0s;
- evidence that the leak survives what FL adds (local SGD, momentum, weight decay, batches, clipping, dropout) and training, with a noise threshold predicted by a singular value;

- an account of what does stop it: local Adam, pruning, noise, narrow or non-overlapping first layers.

2 The span leak

Setting. A client holds features $X = (x_1, \dots, x_T)$, $x_t \in \mathbb{R}^F$. The first layer reads windows $c_p(X) = [x_{ps-a}, \dots, x_{ps-a+k-1}] \in \mathbb{R}^{kF}$, $p = 1..P$, with zeros outside $1..T$, and outputs $Wc_p + b$, $W \in \mathbb{R}^{H \times kF}$. DeepSpeech-1 splices c context frames on each side ($k = 2c + 1$, $s = 1$); the Transformer encoder of [3] uses a convolution with $k = 7$, $s = 3$ on $F = 80$ log-mel bins. The server knows the model and receives the update.

Span. For any loss L and any later layers, with $g_p = \partial L / \partial (Wc_p + b)$,

$$\nabla_W L = \sum_{p=1}^P g_p c_p^\top, \quad \nabla_b L = \sum_{p=1}^P g_p.$$

The row space of $M = [\nabla_W L \mid \nabla_b L]$ lies in $\mathcal{S} = \text{span}\{[c_p, 1]\}$ and equals it when the g_p are linearly independent, which needs $H \geq P$ and holds generically after a nonlinearity. Then $\text{rank } M$ is the number of independent windows: *the update reveals the length of the utterance*.

Linear system. Let $\Pi^\perp$ project onto $\mathcal{S}^\perp$, obtained from the SVD of M . The true features satisfy $\Pi^\perp[c_p(X), 1] = 0$ for all p . Since c_p is linear in X , this is a linear system; its normal matrix is block-banded with bandwidth kF , and we solve it by banded Cholesky. Counting gives a necessary condition for uniqueness: $P(kF + 1 - P)$ equations for about sPF unknowns, hence

$$P \leq (k - s)F + 1.$$

Without overlap ($k = s$) every window may be any vector of $\mathcal{S}$ and nothing is identified.

Local training. After K local SGD steps the server computes $(W_0 - W_K)/\eta = \sum_j \sum_p g_p^{(j)} c_p^\top$. The windows do not depend on the weights, so the row space is the same $\mathcal{S}$. Momentum reweights the terms; weight decay adds a known multiple of W_0 that the server removes. A batch adds the windows of all its utterances to one span. Clipping rescales M . Dropout placed after the first layer changes g_p only.

Noise. If the update is $M + E$ with i.i.d. Gaussian E of standard deviation σ , $\|E\|_2 \approx \sigma(\sqrt{H} + \sqrt{kF + 1})$, and by Wedin's theorem the span is perturbed by an angle of order $\rho = \|E\|_2 / s_P(M)$, with s_P the weakest signal singular value. The attack should degrade linearly in ρ and fail near $\rho = 1$.

3 Beyond the first layer

Past the capacity the first-layer span fills its space. The next layers still leak. In DeepSpeech-1 the second and third layers and the LSTM input are linear maps

Target	Attack	n	Length (s)	Recovered	SNR (dB)
DeepSpeech-1	closed form	1,417	1.3–6.2	98.5%	115
DeepSpeech-1	sequential	191	6.3–35.0	96.9%	127
DeepSpeech-1, trained	closed form	60	≤ 6.2	98.3%	
+ dropout 0.2	closed form	60	≤ 6.2	98.3%	
Transformer-CTC	closed form	600	≤ 9.6	94.8%	63
Transformer-CTC, trained	closed form [†]	60	≤ 9.6	90%	
Gradient matching [12]	2,000 steps [‡]	4	2.3	0%	<3

Table 1: One update per utterance. Median SNR of the recovered features. [†]length given (83% without). [‡]transcript and length given; 2,973 s per utterance against 2.5 s for the closed form.

of $h_1(c_t)$, $h_2(c_t)$, $h_3(c_t)$, clipped-ReLU functions of one window with 2,048 to 4,096 units, and their gradients have row space $\text{span}\{[h_\ell(c_t), 1]\}$ of dimension T for T up to the layer width. The nonlinearity is what raises the capacity: a linear network would leak the same 339-dimensional span at every layer.

A candidate window c can be tested: the residual $r(c) = \Pi_\ell^\perp[h_\ell(c), 1]$ vanishes on true windows, and generically only there, since a kF -dimensional manifold meets a T -dimensional subspace of a much larger space in isolated points. The test does not say *which* frame a window is, so descending on all frames jointly stalls. The padding breaks the symmetry. The first window holds c zero frames followed by $x_1, \dots, x_{c+1}$: we solve for it by Levenberg–Marquardt, starting from the opening frames of public utterances. Every later window shares all but one frame with its predecessor, so each step has only F unknowns. We use the spans of the second and third layers (the LSTM-input span is noisier and is needed only beyond 37 s), re-fit the last 26 frames jointly every 8 steps to stop drift, and stop when six consecutive decoded frames are zero, which is the padding at the far end. The length is therefore an output of the attack. A final joint refinement polishes all frames.

4 Experiments

Setup. LibriSpeech [17] test-clean, one float32 gradient of the CTC loss per utterance. DeepSpeech-1: 26 MFCCs at 50 Hz, $c = 6$, 2,048 units, clipped ReLU. Transformer-CTC: 80 log-mel bins at 100 Hz, per-utterance normalisation, a 1-D convolution from 80 to 1,536 channels ($k = 7$, $s = 3$) with a gated linear unit, 12 blocks. Both are attacked at random initialisation and at checkpoints of training on 100 h. The attacker receives the update and the model only. An utterance counts as recovered if the mean absolute error (MAE) is below 10^{-2} (MFCCs have mean magnitude 8) or the SNR exceeds 30 dB (log-mel).

Closed form. Table 1 and Fig. 1. On DeepSpeech-1 the attack recovers 98.5% of the 1,417 utterances that fit in the first layer, with median MAE 1.4×10^{-5} , and reads the length exactly from the rank. All 21 failures are flagged by a uniqueness test in the solver: these recordings contain runs of bit-identical frames (digital silence), which make windows identical and the solution non-

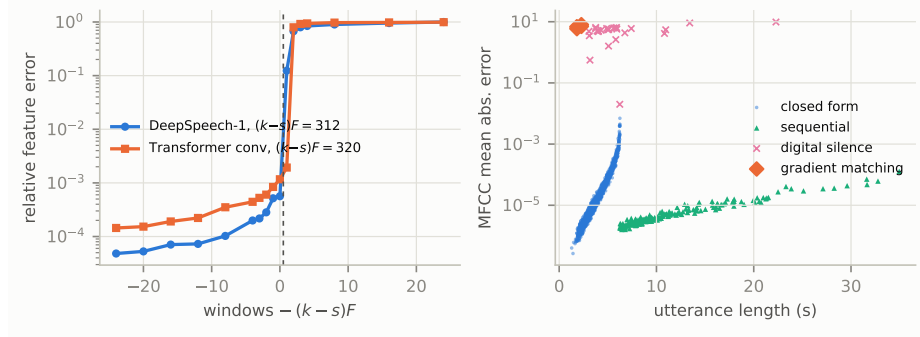


Fig. 1: Left: error of the closed form around the predicted capacity (real features cropped to an exact length). Right: error against utterance length for every attacked utterance.

unique. On the Transformer front end 94.8% of 600 utterances up to 9.6 s are recovered. The capacity is sharp: on four front ends the error stays below 10^{-3} up to $(k-s)F$ windows (312, 468, 320, 240) and reaches order one within two windows.

Long utterances. The sequential decoder recovers 96.9% of 191 utterances between 6.3 and 35.0 s (median MAE 4.0×10^{-6}), including the longest utterance of the set, 1,748 frames. The failures are again digital silence. Trained checkpoints decode equally well (12 of 12).

Baseline. Gradient matching on the output layer, through a CTC loss that can be differentiated twice and with transcript and length given, ends at MAE 5.9 to 8.0 after 2,973 s on 2 s utterances: the gradient is matched (cosine distance below 10^{-2}) and the features are not, as the output-layer gradient has rank at most 29.

What the server receives. With the first-layer update computed from weight differences, the closed form is unaffected by 1 to 50 local SGD steps, momentum and weight decay (median MAE between 10^{-4} and $6 \cdot 10^{-3}$; the floor is float32 rounding in the subtraction and shrinks with larger steps). Batches of 2 to 16 utterances within capacity are recovered when their lengths are given. Dropout up to 0.5 on the client leaves 98% recovered, which contradicts the protection reported for optimisation attacks [6]. Training does not help: DeepSpeech-1 checkpoints from step 0 to 41% word error rate are all recovered at 98.3%.

Audio. A small network maps recovered features to the mel input of a public HiFi-GAN [18], trained on speakers disjoint from the test set. Whisper [19] transcribes the result at 3.8% word error rate (original audio: 3.0%), and ECAPA embeddings [20] identify the speaker among 40 for 98% of reconstructions.

What stops it. Local Adam (per-coordinate rescaling), magnitude pruning of half the entries and sign compression destroy the low-rank structure (0% recovered). Gaussian noise behaves as predicted: the error is proportional to

ρ and the attack fails near $\rho = 1$, which for these utterances is a noise level near 10^{-3} of the gradient RMS. First layers that are narrow or have no overlap do not leak a usable span: we measured the gradient rank of the first layers of pretrained Whisper and wav2vec 2.0 [21], DeepSpeech-2 [22] and QuartzNet [23]; all are saturated. Some leak one layer later (wav2vec 2.0 feature projection, the linear layer after Conformer subsampling [24]), where a decoder analogous to ours finds true windows but recovers only 9% of the frames on average (16 dB) before its chains stall; we count that as a negative result for now. Dropout also stops the sequential decoder, since deeper layers see the dropped activations.

5 Conclusion

A wide linear layer over overlapping feature windows makes one federated update equivalent to the utterance, up to a capacity that we can write down. The defences that work are the ones that break the low-rank structure or exceed the capacity; dropout, clipping, local SGD and training do neither. Central differential privacy adds noise after aggregation and does not protect a client from the server itself. Code, per-utterance results and audio: github.com/takakhoo/unssplice.

References

- [1] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017. arXiv:1602.05629.
- [2] 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)*, 2021. arXiv:2010.15965.
- [3] Martin Pelikan, Sheikh Shams Azam, Vitaly Feldman, Jan Silovsky, Kunal Talwar, Christopher G. Brinton, and Tatiana Likhomanenko. Enabling differentially private federated learning for speech recognition: Benchmarks, adaptive optimizers and gradient clipping. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2310.00098.
- [4] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019. arXiv:1906.08935.
- [5] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients – how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, 2020. arXiv:2003.14053.
- [6] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. arXiv:2104.07815.
- [7] Awni Hannun et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.
- [8] Zhuohang Li, Jiaxin Zhang, and Jian Liu. Speech privacy leakage from shared gradients in distributed learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. arXiv:2302.10441.
- [9] Pretom Roy Ovi and Aryya Gangopadhyay. Gradient inversion attacks on acoustic signals: Revealing security risks in audio recognition systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4835–4839, 2024.

- [10] Xijie Zeng and Frank Rudzicz. How to recover long audio sequences through gradient inversion attack with dynamic segment-based reconstruction. In *Proc. Interspeech*, pages 5118–5122, 2025.
- [11] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning (ICML)*, 2006.
- [12] Minh N. Bui, Trung Dang, Paul Cherian, Trac D. Tran, and Peter Chin. Reconstructing speech features of automatic speech recognition systems in federated learning by gradient descent. In *Complex Networks & Their Applications XIII (COMPLEX NETWORKS 2024)*, Studies in Computational Intelligence, pages 125–134. Springer Nature Switzerland, 2025.
- [13] Jiacheng Du, Jiahui Hu, Zhibo Wang, Peng Sun, Neil Zhenqiang Gong, Kui Ren, and Chun Chen. SoK: On gradient leakage in federated learning. In *USENIX Security Symposium*, 2025. arXiv:2404.05403.
- [14] Ivo Petrov, Dimitar I. Dimitrov, Maximilian Baader, Mark Niklas Müller, and Martin Vechev. DAGER: Exact gradient inversion for large language models. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2405.15586.
- [15] Dimitar I. Dimitrov, Maximilian Baader, Mark Niklas Müller, and Martin Vechev. SPEAR: Exact gradient inversion of batches in federated learning. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2403.03945.
- [16] William Kalikman, Ivo Petrov, Dimitar I. Dimitrov, and Martin Vechev. TIGER: Inverting transformer gradients via embedding-subspace distance optimization. *arXiv preprint arXiv:2606.18312*, 2026.
- [17] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- [18] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, 2020. arXiv:2010.05646.
- [19] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, PMLR 202, 2023. arXiv:2212.04356.
- [20] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. Interspeech*, 2020.
- [21] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.
- [22] Dario Amodei et al. Deep speech 2: End-to-end speech recognition in english and mandarin. *arXiv preprint arXiv:1512.02595*, 2015.
- [23] Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. QuartzNet: Deep automatic speech recognition with 1D time-channel separable convolutions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
- [24] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.