Long-Form Speech Reconstruction from Gradients in Federated ASR using CTC Loss

Taka Khoo and Minh N. Bui and Peter Chin

Dartmouth College
Hanover NH, United States

Abstract. Federated Learning (FL) is increasingly used in Automatic Speech Recognition (ASR) to protect user privacy. However, recent work demonstrates that gradients can leak sensitive information. Existing speech reconstruction attacks typically avoid Connectionist Temporal Classification (CTC) loss due to its non-aligned structure and training instability, limiting prior approaches to simplified or short-utterance settings. As a result, methods that bypass CTC can only be applied to speech *classification* tasks or toy ASR scenarios, rather than full end-to-end speech recognition. In contrast, we show that CTC *enables* high-fidelity gradient inversion for realistic end-to-end ASR models. We introduce a stable optimization framework capable of reconstructing long-form speech (10–30s) from FL gradients using CTC-aware inversion, revealing severe privacy leakage in practical ASR deployments.

1 Introduction

Federated Learning (FL) enables clients to collaboratively train models without sharing raw data, making it appealing for ASR [1, 2, 3] where audio is highly sensitive. However, gradients exchanged during FL can leak information about inputs [4], as shown by image reconstruction attacks in computer vision [5].

For speech, prior works demonstrate gradient-based audio reconstruction [6, 7, 8], but mostly avoid the CTC loss [9] or focus on short segments. CTC introduces alignment complexity, repeated-token collapse, and long-range dependencies, making long-utterance inversion challenging.

We propose a CTC-compatible gradient inversion framework that stabilizes optimization for utterances up to 10 seconds, enabling recovery of both semantic content and speaker characteristics.

Contributions:

- We present the first demonstration of long-form speech reconstruction from FL gradients *using* CTC loss.
- We introduce a stable optimization pipeline tailored for CTC-based models, incorporating multi-resolution losses, length-scaling, and chunked alignment refinement.

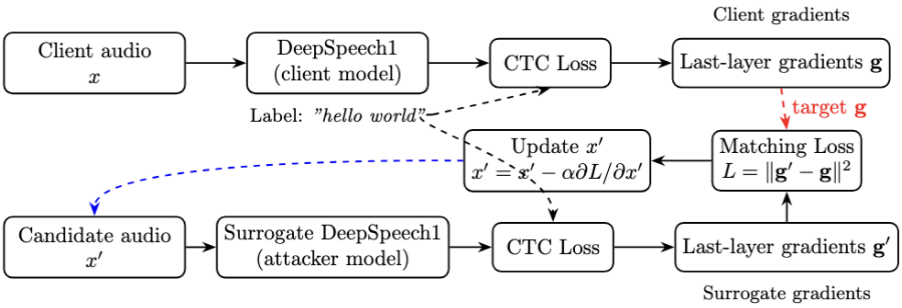

Fig. 1: Overview of our gradient matching pipeline.

2 Related Work

2.1 Gradient Leakage Attacks

Gradient inversion [4] recovers a private training sample $\mathbf{x}$ from the shared gradient $g = \nabla_\theta \mathcal{L}(\mathbf{x}, y)$ by optimizing a synthetic input $\hat{\mathbf{x}}$ to match g. Early speech extensions impose simplifying assumptionsâshort audio windows, fully-aligned losses, or proxy teacher–student modelsâto bypass the non-aligned CTC objective, limiting attacks on long-form or production ASR.

2.2 ASR Models and CTC Loss

End-to-end ASR systems like *DeepSpeech-1* [10] use CTC to handle unaligned transcripts. CTC introduces non-smooth, alignment-dependent gradients, complicating reconstruction. Prior works circumvent this with frame-aligned cross-entropy [8], simplified settings [7], or costly optimization [6].

CTC-based architectures remain common in resource-constrained federated learning [11]. We therefore focus on DeepSpeech-1 and show that full CTC still allows substantial gradient leakage, challenging the assumption that alignment uncertainty protects privacy.

3 Methods

3.1 Base Gradient-Matching Attack

We follow the DeepSpeech gradient inversion attack of [6] (Figure 1). A client computes the CTC loss $L_{\mathrm{CTC}}(X, y)$ on MFCCs X and returns its gradient $g^\star$. The attacker mirrors the model, initializes a dummy MFCC $\hat{X}$, and matches gradients by minimizing a cosine meta-loss $\mathcal{D}(\hat{X}, g^\star)$.

Full algorithmic details are given in [7], we focus only on the extensions required for long-utterance reconstruction-namely grid-based MFCC reconstruction, grid-aware learning rates, and alternating grid updates.

3.2 Log-space CTC for numerical stability

When extending to longer sequences, if forward–backward probabilities are accumulated in probability space, the CTC loss quickly overflows to ∞ and causes gradient vanishing.

To stabilize training, we thus switch to a log-space CTC loss. Given logits $z_{t,c}$ for character c at frame t, we compute log-probabilities via log-softmax and run the entire CTC recursion in the log domain. We denote the resulting loss by $L_{\mathrm{CTC}}^{\log}(X, y)$ and use it on both the client $g^\star = \nabla_\theta L_{\mathrm{CTC}}^{\log}(X, y)$ and attacker sides.

3.3 Grid-based reconstruction

Long utterances produce MFCCs $X \in \mathbb{R}^{T\times F}$ with large T, and we found that directly optimizing a full-length dummy $\hat{X}$ yields visually unstable reconstructions for 8–10 second inputs. Although framewise metrics remain usable, the resulting MFCCs lack coherent temporal structure. To address this, we reparameterize $\hat{X}$ through a set of overlapping *grids* that break the sequence into shorter, more stable segments.

Given a grid width W (frames) and stride $S \le W$, we segment X along the time axis into

$$X \;\Rightarrow\; \{X^{(g)}\}_{g=1}^{G}, \qquad X^{(g)} \in \mathbb{R}^{W\times F},$$

where grids overlap when $S < W$. A corresponding dummy set $\{\hat{X}^{(g)}\}$ is optimized by the attacker, and concatenating these dummy grids, averaging any overlapping regions, yields a full-length $\hat{X}_{\mathrm{full}}$ compatible with the DeepSpeech model. The gradient-matching loss remains unchanged.

3.4 Sequential Grid Reconstruction with Per-Grid Learning-Rate Resets

We first optimize grids in fixed temporal order, updating only the selected grid $\hat{X}^{(g)}$ at each step. Using a single global learning rate causes later grids to collapse, as their updates occur when the LR has decayed.

To fix this, we reset each grid's LR and Adam state when moving to a new grid, applying a simple decay schedule $\eta_{g,k_g} = \eta_0\,\gamma^{\lfloor k_g / K_{\mathrm{decay}}\rfloor}$, $0 < \gamma < 1$. This ensures every grid, including late ones, receives sufficiently high LR, making sequential reconstruction viable for long MFCC sequences.

3.5 Alternating grid schedules

Sequential updates refine each grid fully before moving on, but long utterances contain regions of uneven difficulty. Inspired by segment-alternation strategies [8], we treat grid selection as a scheduling problem.

At iteration t, we pick a grid $g_t \in \{1, \ldots, G\}$, rebuild $\hat{X}_{\mathrm{full}}$, compute $\mathcal{D}$, and update only that grid with its per-grid LR. We consider: (i) *Round-robin*, cycling through grids sequentially, and (ii) *Error-driven*, repeatedly updating the

grid with the worst MAE. Hybrid schedules alternate between these strategies, allowing difficult segments to receive repeated high-LR updates and overcoming bottlenecks of strict left-to-right optimization.

4 Experiments

4.1 Setup and Metrics

We attack 10s LibriSpeech utterances with the original model, log-space CTC, and the reconstruction pipeline from Section 3. Optimization (Adam, 0.5 LR, MultiStep halving every 250 steps, 10 seeds per grid) is fixed; only grid geometry, update schedule, and blending/settle passes vary. We report: (i) cosine gradient misalignment $\mathcal{D}$, tracking seam energy; (ii) global MFCC MAE, inflated by overlap revisits; and (iii) waveform SNR after inverse MFCC synthesis. MAE/SNR favor smoothed overlaps, so we interpret them alongside Figure 2.

4.2 Progression of Reconstruction Regimes

We evaluate four configurations with increasing structure:

Seq. 300×150. Single-pass grids with per-grid LR resets prevent tail collapse but degrade the head; MAE is moderate (8.78) due to minimal early-grid overlap.

Alt. 300×150. Cycling grids avoids head collapse, but large overlaps average seams in opposite directions ($\mathcal{D} = 5.31{\times}10^{-3}$, SNR $= -1.52$ dB).

Alt. 288×96. Reduced grid width/overlap and aligned stride remove remainder tiles, halving cosine loss ($2.76{\times}10^{-3}$) while raising MAE to 10.74 from more boundary revisits.

Alt. 288×96 + priority MAE. Error-driven scheduling, normalized blending, and a 100-step settle pass produce the best reconstructions: $\mathcal{D} = 1.64{\times}10^{-3}$, MAE $= 9.81$, SNR $= -1.63$ dB.

Experiment	$\mathcal{D}$	MAE	SNR (dB)
Seq. 300/150	5.63×10^{-3}	8.78	-0.33
Alt. 300/150	5.31×10^{-3}	8.99	-1.52
Alt. 288/96	2.76×10^{-3}	10.74	-2.04
Alt. 288/96 (err-driven)	$\mathbf{1.64 \times 10^{-3}}$	**9.81**	-1.63

Table 1: Average metrics across six utterances (original sample + two additional clips per regime).

4.3 Metrics and Cross-Sample Consistency

Averaging over six LibriSpeech utterances preserves the ordering in Table 1: sequential 300/150 is always worst, cyclic 300/150 stabilizes only the head, aligned 288/96 removes seam drift, and the priority 288/96 recipe is best. The standard deviation of $\mathcal{D}$ is below 6% of each mean, and MAE variance 0.6 is far smaller

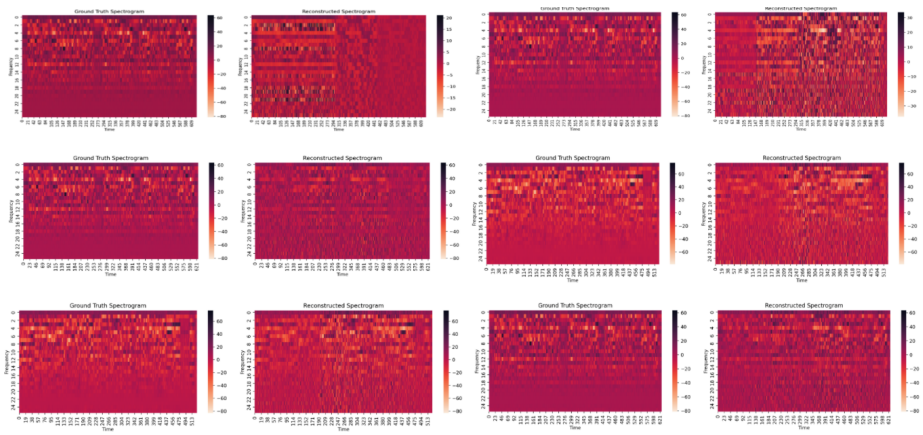

Fig. 2: Progression of reconstructions, shown as paired ground-truth (left) and reconstructed (right) spectrograms. Row 1: sequential grids without/with per-grid LR resets. Row 2: alternating 300/150 and alternating 288/96. Row 3: priority-MAE alternating 288/96 with normalized blending and a settle pass (two utterances).

than the 1.5-2.0 gaps between regimes, confirming that the improvements stem from the design changes rather than speaker-specific quirks.

4.4 Qualitative Evaluation

Figure 2 summarizes the visual story. Row 1 shows sequential baselines: global schedules collapse the head, while per-grid resets fix the tail but leave early frames blurry. Row 2 uses alternating updates: the 300/150 layout reduces collapse but leaves seam ghosts, whereas the aligned 288/96 layout eliminates seam drift despite higher MAE. Row 3 applies the full recipe; seams disappear, fricatives remain bright, and reconstructions closely match the ground truth. High-frequency bands are slightly sharper due to CTC gradients, but intelligibility is preserved. These trends align with metric improvements, showing that alternating schedules, aligned grids, and blending passes jointly address collapse, seam drift, and smoothing.

5 Conclusion

We show that long-form speech in Federated ASR is vulnerable to gradient leakage even when trained with CTC loss. Our CTC-aware reconstruction framework reveals that alignment information embedded in gradients enables high-fidelity inversion. These findings call for stronger privacy defenses in real-world FL-ASR systems.

References

[1] X. Cui, S. Lu, and B. Kingsbury, "Federated acoustic modeling for automatic speech recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6748–6752, 2021.

[2] D. Dimitriadis, K. Kumatani, R. Gmyr, Y. Gaur, and S. E. Eskimez, "A federated approach in training acoustic models," in *Interspeech 2020*, pp. 981–985, 2020.

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

[4] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," 2019.

[5] H. Yin, A. Mallya, A. Vahdat, J. M. Alvarez, J. Kautz, and P. Molchanov, "See through gradients: Image batch recovery via gradinversion," 2021.

[6] T. Dang, O. Thakkar, S. Ramaswamy, R. Mathews, P. Chin, and F. Beaufays, "A method to reveal speaker identity in distributed asr training, and how to counter it," 2021.

[7] M. N. Bui, T. Dang, P. Cherian, T. D. Tran, and P. Chin, "Reconstructing speech features ofÂ automatic speech recognition systems inÂ federated learning byÂ gradient descent," in *Complex Networks & Their Applications XIII* (H. Cherifi, M. Donduran, L. M. Rocha, C. Cherifi, and O. Varol, eds.), (Cham), pp. 125–134, Springer Nature Switzerland, 2025.

[8] X. Zeng and F. Rudzicz, "How to recover long audio sequences through gradient inversion attack with dynamic segment-based reconstruction," pp. 5118–5122, 08 2025.

[9] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, (New York, NY, USA), p. 369â376, Association for Computing Machinery, 2006.

[10] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, vol. abs/1412.5567, 2014.

[11] A. Agarwal and T. Zesch, "Ltl-ude at low-resource speech-to-text shared task: Investigating mozilla deepspeech in a low-resource setting," 06 2020.