RECONSTRUCTION OF PRIVATE SPEECH USING GENERATIVE-PRIOR-GUIDED GRADIENT INVERSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning enables collaborative training without centralising raw audio, yet the gradients shared during training can leak sensitive information, including speech content and speaker identity. Existing audio gradient inversion attacks primarily reconstruct inputs directly in the feature space and have been studied mostly under relatively favourable inversion settings. In this work, we introduce a generative-prior gradient inversion framework for audio, which constrains reconstruction using a learnt mel-spectrogram prior and employs a neural vocoder for waveform recovery. To prevent prior memorisation of victim identities, the generator is trained using an identity-disjoint speaker split. We evaluate the attack on Speech Commands V2 and AudioMNIST under a range of increasingly challenging settings, including gradient clipping with Gaussian perturbation, compressed MFCC features, trained victim models, and different network architectures. Our results show that the main advantage of the generative prior emerges as the inversion problem becomes less informative. Under the strongest tested gradient perturbation, speaker recovery remains at 84.0% on AudioMNIST and 96.5% on Speech Commands, compared with 18.0% and 29.3% for the GAN-free baseline. These findings demonstrate that degraded speech reconstruction does not ensure speaker privacy, as identity leakage can persist through generative priors.

1 INTRODUCTION

Federated Learning (FL) (McMahan et al., 2017) is being widely adopted, extending across modalities, use-cases, and domains. FL primarily extends Machine Learning (ML) across privacy barriers by allowing models to train on broader datasets without requiring access to private data, aggregating local gradients into a global model. The private nature benefits sensitive domains such as healthcare, voice assistants, and information rich modalities such as audio. For instance, FL can be used for disease detection tasks, such as Parkinson's disease detection using patient speech data (Tayebi Arasteh et al., 2023; Tayebi Arasteh et al., 2025). The audio modality is unique compared to other modalities, as it carries unique details: speech content, speaker identity & attributes. At the same time, these systems cannot be centralised due to regulatory requirements (GDPR (EU, 2016), HIPAA (US Congress, 1996)), making FL a viable alternative in sensitive domains.

Deep Leakage of Gradients (DLG) (Zhu et al., 2019) introduced Gradient Inversion (GI), where a curious server recovers the private data used in FL training by solving the gradient matching objective on the gradients shared by the clients. Since then, Gradient Inversion Attacks (GIA) have evolved rapidly. For instance, image GIA has matured with GAN-prior manifold attacks (Li et al., 2022; Jeon et al., 2021; Fang et al., 2023), and several architectural variants have been established as well (Hatamizadeh et al., 2022). Similarly, text GIA is widely explored with modality specific improvements (Gupta et al., 2022; Balunović et al., 2022). However, despite carrying more sensitive details, the audio GIA remains under-explored. Prior work infers speaker identity from recovered ASR features (Dang et al., 2022), adapts DLG to recover waveforms (Li et al., 2023), and extends it to long audio (Zeng & Rudzicz, 2025). These attacks do not utilise GAN priors, and identity leakage is suppressed where gradient is defended (Dang et al., 2022). It remains unclear whether a learnt audio prior sustains leakage, particularly speaker identity, when gradient information is degraded.

In this work, we study this question with a generative-prior guided GIA for audio. Analogous to GAN-based image GIA (Fang et al., 2023), we constrain the search with a modified StyleGAN2

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Karras et al., 2020) prior over mel-spectrograms, and recover waveforms with the BigVGAN v2 vocoder[1] (Lee et al., 2023) instead of the Griffin-Lim (GL) algorithm used in prior audio GI (Li et al., 2023; Zeng & Rudzicz, 2025). We evaluate the attack as the inversion problem becomes information-limited: under gradient perturbation, compressed Mel-Frequency Cepstral Coefficient (MFCC) features, trained victims, and different architectures. We find that the generative prior provides its largest advantage when the inversion problem becomes ill-posed. Under gradient perturbation and compressed MFCC features, the proposed attack retains substantially more speaker identity than the GAN-free baseline. The gains are architecture dependent: substantial leakage persists for convolutional victims, whereas ResNet18 and AST remain considerably more resistant. An important consequence is that reduced content intelligibility should not be interpreted as evidence of speaker privacy, as the generative attack retains identity-relevant structure even when conventional reconstruction degrades substantially. Additionally, we verify the utility of general priors in conducting cross-dataset attacks and empirically relax the strong attack side assumptions.

The key contributions of this work can be listed as follows:

- We propose a generative-prior gradient inversion attack for audio that constrains feature recovery to a mel-spectrogram GAN manifold and recovers waveforms with a neural vocoder.
- We quantify the gain from a pre-trained neural vocoder (BigVGAN v2) over the conventional Griffin-Lim algorithm.
- We show that, under gradient perturbation and MFCC compression, the prior preserves speaker identity at levels where the GAN-free attack fails, even as fidelity metrics degrade.
- We characterise the limits of this effect across trained victims and architectures, and show that poor content reconstruction alone is not sufficient evidence of speaker privacy.

2 Background and related work

DLG (Zhu et al., 2019) first introduced the FL vulnerability to Data Reconstruction Attacks (DRA) as an optimisation problem with the gradient matching objective, primarily for the image modality. Subsequent work (Geiping et al., 2020; Yin et al., 2021; Zhao et al., 2020) extended the optimisation with improved objectives, regularizations, and label recovery. While the early methods operated in the input space, another class of GAN-based attacks (Li et al., 2022; Jeon et al., 2021; Fang et al., 2023) emerged that utilises latent priors to constrain the search space. Particularly, GIFD (Fang et al., 2023) is a recent SOTA image gradient inversion method operating in the feature domain via Intermediate Layer Optimisation (ILO) (Daras et al., 2021). GIFD uses generators such as StyleGAN2 (Karras et al., 2020) and BigGAN (Brock et al., 2019) to generate the candidate images for the gradient matching objective. StyleGAN2 exposes an intermediate latent space vector that can be used to generate images that match the original data. We inherit this class of GAN-based attacks from the image domain to implement the proposed Generative GIA framework in the audio domain.

Speech Privacy Leakage (SPL) (Li et al., 2023) and Hessian Free Gradients Matching (HFGM) (Dang et al., 2022) on Automatic Speech Recognition (ASR) models formalised two staged GIAs for the audio, with SPL treating audio features as images. The first stage of both works solves the gradient matching objective to recover the audio features in the input space. SPL uses Anisotropic Total Variation (ATV) (Lou et al., 2015) as a regularizer and in the second stage it uses the Griffin-Lim (GL) (Griffin & Lim, 1984) algorithm to recover the waveform. In the second stage HFGM conducts speaker identification, but is limited to MFCCs and the ASR model. Also, HFGM shows that it degrades under gradient perturbations. Subsequently, to extend the gradient inversion attack to long audio sequences (LAS) (Zeng & Rudzicz, 2025), a segmented approach was proposed. While these methods measure the identity leakage aspect, unique to audio data, they do not utilise GAN priors. Additionally, both SPL and LAS methods use the GL algorithm as the vocoder and HFGM method does not recover the waveform. However, generative vocoders such as BigVGAN v2 (BVG) (Lee et al., 2023) reconstruct the lost phase much better. App. A.1 details these and related attacks.

In contrast to the image modality, the audio modality is vulnerable in multiple aspects; speaker identity, content, attributes, and fidelity. For instance, in the case of FL use for Parkinson's Disease diagnosis (Tayebi Arasteh et al., 2023; Tayebi Arasteh et al., 2025), a successful attack could leak

[1] https://huggingface.co/nvidia/bigvgan_v2_22khz_80band_fmax8k_256x

the vocal identity of the patient, the content of the patient utterance, attributes such as age, gender, accents, and the bio-markers. Therefore, we use various metrics to measure the leakage along these different axes (Taal et al., 2010; Rix et al., 2001; Desplanques et al., 2020) (Sec. 4.3).

Differential privacy (DP-SGD) (Abadi et al., 2016) has been considered the gold standard in Privacy Preserving ML (PPML) for a long time. It protects the gradient by clipping and adding Gaussian noise. In contrast, the recent SOTA defence Censor (Zhang et al., 2025a) uses an orthogonal gradient approach to prevent leakages. Given its wide adoption and a single parameter controlling perturbation strength, we use the clipping and Gaussian-noise mechanism of DP-SGD (Abadi et al., 2016) as our main gradient perturbation (Sec. 5.1) and additionally evaluate Censor (App. A.6).

3 METHODOLOGY

In this section, we formalise the FL configuration and the Gradient Inversion setting. Then we discuss the existing gradient inversion context leveraged as a foundation. Finally, we elaborate on the methodology proposed to extend generative gradient inversion to the audio modality.

3.1 PROBLEM FORMULATION

We consider a standard distributed learning setup where clients $c \in C$ train a local model on private data D_c and sends gradients $\nabla_\theta L$ to a centralised server to aggregate it into a global model. To handle the audio modality, the training data is transformed into audio features using $\mathcal{T}_{mel}(x_i) = u_i$ and u_i is input to the model F. The local model has a classification objective; $\min_\theta \forall_{D_c} L(F(u_i), y_i)$, where $(x_i, y_i) \in D_c$ are the training data, and $g = \nabla_\theta L(F(u_i), y_i)$ is the shared gradient.

Threat Model: As the threat model, we assume an honest-but-curious server observing the client gradient g_{obs} that does not interfere with the training. Unless stated otherwise, we consider the base case of an untrained victim model sharing a single sample gradient with the server for the attack. Consistent with previous work (Zhu et al., 2019; Zhao et al., 2020; Geiping et al., 2020), we assume the curious server has knowledge of the model architecture, global model parameters, labels, and the ability to store the shared gradients and conduct the attack offline if needed. The curious server is aware of the clip length and the mel-spectrogram configuration used to train the model. Similar to the generative GIA work in the image domain (Jeon et al., 2021; Fang et al., 2023; Li et al., 2022), the attacker has access to a generative prior G trained on a similar data distribution as the clients. However, the generative prior is trained on an id-disjoint split; victim client speakers are omitted in the training data of the prior. Therefore, the prior cannot memorise the victim speakers in advance. The server is not aware of the perturbation $\mathcal{P}$ applied to the gradient ($g_{\text{obs}} = \mathcal{P}(g)$, App. A.6).

3.2 FOUNDATION

The main objective of optimisation based GI (Zhu et al., 2019) is to solve the gradient matching problem adapted as described in Eq. 3. The gradient matching loss $\mathcal{L}_{\text{grad}}$ is calculated by the distance $\mathcal{D}$ between the observed gradient g_{obs} and the dummy gradient $\hat{g} = \nabla_\theta L(F(\hat{u}_i), y_i)$. The choice of distance functions has been studied in previous work (Zhu et al., 2019; Geiping et al., 2020) and we empirically choose a weighted combination of the $L2$ and $Cosine$ distances.

$$\mathcal{L}_{\text{grad}} = \mathcal{D}(g_{\text{obs}}, \hat{g}) \tag{1}$$

Similar to the usage of Total Variation (TV) (Rudin et al., 1992) in the image domain, Audio GI uses Anisotropic TV (ATV) (Li et al., 2023) as a loss L_{ATV} to handle regularization, but smoothing is kept lower across the frequency axis ($\lambda_h > \lambda_v$).

$$\mathcal{L}_{\text{ATV}} = \lambda_h \, \text{TV}_h(\hat{u}_i) + \lambda_v \, \text{TV}_v(\hat{u}_i), \qquad \lambda_h > \lambda_v \tag{2}$$

3.3 GENERATIVE AUDIO GRADIENT INVERSION

In this section, we discuss the Generative Audio GI approach step by step (Alg. 1). The methodology is broken down into the following phases; 1) Mel-spectrogram GAN prior, 2) Feature recovery to

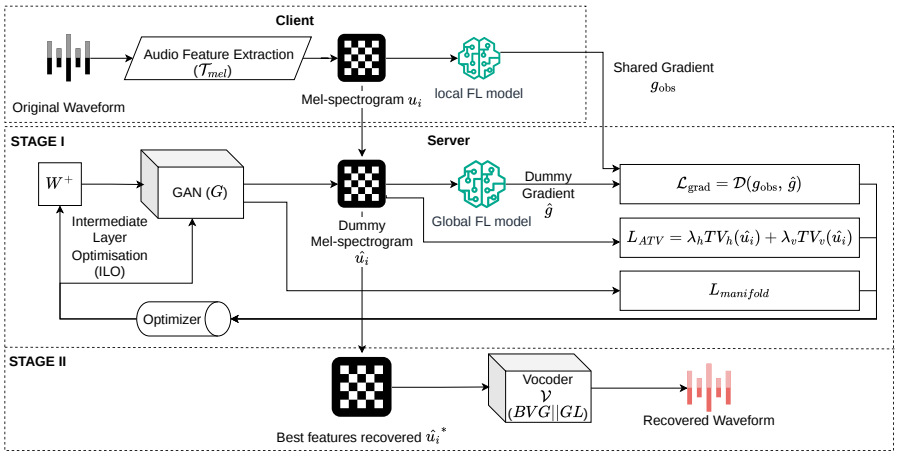


Figure 1: Overview of the two-stage attack: Stage I searches W^+ and then ILO to match dummy gradient $\hat{g}$ to the observed gradient g_{obs}. Stage II vocodes the best spectrogram $\hat{u}_i^*$ into a waveform.

approximate the audio feature input vector, 3) Waveform recovery by converting the recovered audio features back to speech. Note that due to the offline attack capability, these steps can be conducted with sufficient time and compute at the cost of privacy for the client. An overview of the attack methodology is depicted in Fig. 1.

Mel-spectrogram GAN prior: We adopt a single-channel StyleGAN2 (Karras et al., 2020) like GAN as our prior architecture. The choice of this architecture is derived from the GIFD (Fang et al., 2023) image domain attack to leverage fine-grained control of the features with ILO (Daras et al., 2021). The intermediate layer conditions are controlled by the style vector W. The extended latent space denoted by W^+ is used with a $w \in W$ for each layer for granular conditioning. Furthermore, we use valid audio augmentations like pitch shifting at data preparation time and non-leaking audio augmentations (translation and cutout) at training time to handle the low volume of datasets. All mel-spectrograms use a fixed configuration (App. A.5) kept uniform across data preparation, GAN outputs, victim model inputs, and vocoders for compatibility and simplicity.

Stage I - Feature Reconstruction Phase: The first stage recovers the audio features using GAN-based optimisation. To create the dummy input features $\hat{u}_i$, the GAN G conditioned on the latent W is used, where $\mathcal{C}$ crops the real region from the GAN canvas (App. A.5) and applies the inverse global affine to return a physical log-mel. The latent vector W is initialised using the average latent w_{avg} inferred at GAN training time to start with a realistic mel $\hat{u}_i$ within the manifold. To handle the local minima in GAN latent space, we use a best-of-N warm start: we probe a pool of N jittered candidates $W_j^+ = w_{\text{avg}} + \epsilon_j$ with $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$, score each by its gradient-matching distance $\mathcal{D}$ against the observed gradient g_{obs}, and retain the top-K as the parallel restarts that are then fully optimised. Then, we minimise the distance $\mathcal{D}$ between g_{obs} and the dummy gradient $\hat{g}$ as per the objective in Eq. 3, where $\phi_{s_0} = W^+$ and $G_{>s_0} = G$ in the initial phase. Each restart is refined under an ILO schedule $\mathcal{S}$: the first phase optimises the extended latent itself ($\phi_{s_0} = W_k^+$), while later phases unlock and optimise the intermediate activation $\phi_s = G_{\leq s}(W_k^+)$, synthesising the mel through the downstream sub-network $G_{>s}$. The active variable ϕ_s is updated to minimise the total loss $\mathcal{L}_{\text{total}}$ (Eq. 4), and the manifold penalty is applied to ϕ_s at each intermediate-layer phase.

$$\phi_s^* = \arg\min_{\phi_s} \mathcal{D}\big(g_{\text{obs}},\ \nabla_\theta L\big(F(\hat{u}_i),\, y_i\big)\big)\,, \qquad \hat{u}_i = \mathcal{C}\big(G_{>s}(\phi_s)\big) \tag{3}$$

In addition to the gradient matching loss L_{grad}, we use the Anisotropic TV (ATV) loss L_{ATV} for smoothing and the $L_{manifold}$ loss consistent with GIFD (Fang et al., 2023) for realistic mel-spectrogram generation in the total loss L_{total} as regularizations (Eq. 4). $L_{manifold}$ is an l_2 penalty to keep the latent near the realistic manifold.

Algorithm 1 Generative audio gradient inversion

Require: Observed gradient $g_{\text{obs}} = \mathcal{P}(g)$; label y_i; victim F_θ; generator G, w_{avg}; downstream sub-networks $G_{>s}$
Require: probe pool N; restarts K; ILO layer schedule $\mathcal{S} = (s_0, \ldots, s_M)$; iters per phase T
Ensure: recovered waveform $\hat{x}_i^*$ and mel spectrogram $\hat{u}_i^*$

```
sample pool {W_j^+ = w_avg + ε_j}_{j=1}^N, ε_j ~ N(0, σ^2)                ▷ probe N candidates
{W_k^+}_{k=1}^K ← top-K_j [−D(g_obs, ∇_θ L(F(C(G(W_j^+))), y_i))]          ▷ best-of-N warm starts
for k = 1 to K do                                                          ▷ parallel restarts
    for phase s in schedule S do                                           ▷ progressive ILO over layers
        unlock layer-s variable φ_s                                        ▷ φ_{s_0} = W_k^+; else φ_s = G_{≤s}(W_k^+)
        for t = 1 to T do
            û_i ← C(G_{>s}(φ_s))                                           ▷ downstream synthesis + crop + inv. affine
            ĝ ← ∇_θ L(F(û_i), y_i)
            L_total ← D(g_obs, ĝ) + L_ATV(û_i) + L_manifold(φ_s)
            φ_s ← Adam-step(∇_{φ_s} L_total)
        end for
    end for
    û^(k) ← C(G_{>s}(φ_s));  m_k ← D(g_obs, ∇_θ L(F(û^(k)), y_i))          ▷ final recon + match
end for
k* ← arg min_k m_k                                                         ▷ min-grad-match selection
û_i* ← û^(k*)
x̂_i* ← V(û_i*)                                                             ▷ vocode once, outside loop
return x̂_i*, û_i*
```

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{grad}} + \mathcal{L}_{\text{ATV}} + \mathcal{L}_{\text{manifold}} \tag{4}$$

Stage II - Waveform Reconstruction Phase: In this stage, the best recovered features $\hat{u_i}^*$ are used to recover the approximated waveform $\hat{x_i}^*$ (Eq. 5). While previous work relies on the Griffin-Lim (GL) algorithm (Griffin & Lim, 1984), we propose the use of the SOTA BVG (Lee et al., 2023) as the second stage vocoder $\mathcal{V}$. This is due to the inconsistent phase reconstruction artefacts introduced by the GL algorithm, compared to the modern BVG vocoder. Also, BVG (Lee et al., 2023) is considered a good zero-shot vocoder for out-of-distribution tasks. The vocoder is used outside of the optimisation loop and is applied to the baseline as well for fairness.

$$\hat{x}_i^* = \mathcal{V}(\hat{u}_i^*), \qquad \hat{u}_i^* = \mathcal{C}\big(G_{>s_M}(\phi_{s_M}^*)\big) \tag{5}$$

3.3.1 Extending to MFCC

In order to make the attack methodology compatible with MFCC features, we extend the existing stages as follows. We use the dummy Mel $\hat{u}_i$ generated by the GAN to produce the dummy MFCC $\hat{v}_i$ using the Discrete Cosine Transform (DCT) function $\mathcal{T}_{DCT}$ as explained by Eq. 6. The victim model consumes the MFCC features in this case, and the $\mathcal{T}_{DCT}$ function is differentiable.

$$\hat{v}_i = \mathcal{T}_{DCT}(\hat{u}_i), \qquad \hat{u}_i = \mathcal{C}\big(G_{>s}(\phi_s)\big) \tag{6}$$

4 Experiments

In this section, we discuss the experimental setting used to evaluate the attack performance of the proposed methodology. We describe the datasets, target model architectures and metrics used to evaluate the performance, and the GAN training details. While we use both BVG (Lee et al., 2023) and GL (Griffin & Lim, 1984) vocoders, we report the superior BVG in the main results in Section 5 and discuss the impact in Section 6. The additional configurations are discussed in App. A.5.

4.1 DATASETS

We use the Speech Commands V2 (SC) (Warden, 2018) and the AudioMNIST (AM) (Becker et al., 2024) datasets to conduct the evaluation of our methodology. Both datasets contain $\sim$1 second audio clips from diverse speakers with labels for classification objectives. The SC dataset has 35 labels with simple natural language commands ('yes', 'no', 'up', 'down', etc.) and digits (0-9), while the AM dataset has only digits (0-9). We consider a randomly chosen fixed set of 100 audio clips from 10 speakers in the AM dataset and 400 audio clips from 200 speakers in the SC dataset. The remaining data is used for three purposes; 1) to train the audio mel-spectrogram GAN prior, 2) to train networks to evaluate the attack at later epochs, and 3) to evaluate model utility under defences.

Id-disjoint split: To avoid the strong assumption that the GAN has seen the victim speakers in advance, the dataset used to train the GAN does not contain any of the speakers targeted in the attack. Therefore, an attacker can use a general prior in the victim data distribution with sufficient speaker diversity to train the prior without having access to the audio clips of the speakers.

4.2 VICTIM MODELS

Consistent with the keyword spotting architecture used in the SPL attack (Li et al., 2023) evaluation, we use the same CNN based model referred to as AudioCNN (App. A.5.1) for which all experiments are conducted on unless stated otherwise. In order to conduct experiments under diverse ill-posed settings, we also use a LeNet (LeCun et al., 1998) style model, ResNet18 (He et al., 2016), and an Audio Spectrogram Transformer (AST) (Gong et al., 2021) style model. The victim model details can be found in App. A.5.1.

4.3 METRICS

We use different metrics to cover each axis of leakage in audio; content, intelligibility, and speaker identity. To measure content leakage, we use mel-spectrogram MSE (Mel-MSE). Although mel-spectrogram cosine similarity (Mel Cos.Sim) has been used in past work (Zeng & Rudzicz, 2025), it is not sufficiently informative (App. A.4). To measure speech intelligibility, we use the standard STOI (Taal et al., 2010) metric (higher is better, [0,1]). PESQ (Rix et al., 2001) (higher is better, [-0.5,4.5]) is also used as an audio quality measure, which is broadly regarded as an industry standard; however, it is relatively less stable for short audio sequences. To evaluate speaker identity leakage, we use the standard ECAPA-TDNN (Desplanques et al., 2020) cosine similarity (ECAPA Cos.Sim). We also use the ECAPA Cos.Sim threshold value $\tau = 0.25$ (App. A.3) to find the percentage of clips of which identity is recovered to measure the speaker ID leakage rate (Speaker ID %). While waveform MSE (W-MSE) and Signal-to-Noise ratio (SNR) can be useful metrics in other contexts, we omit the use of these metrics for several reasons. The inversion process is noisy and introduces artefacts to the mel-spectrogram. Additionally, both vocoding mechanisms (GL and BVG) can introduce phase artefacts while converting mel-spectrograms into waveforms. Hence, we omit the W-MSE and SNR metrics to avoid discrediting intelligible reconstructions.

4.4 GAN TRAINING

We train one GAN prior per each dataset. The AM and SC dataset priors are trained with 300,000 and 400,000 iterations respectively, and we chose the best checkpoint based on the Fréchet distance (FD) (Heusel et al., 2017) and the entropy converging to the expected data distribution of each dataset. Post split, the AM prior is trained on only $\sim$25,000 clips from $\sim$50 speakers, below the recommended amount for StyleGAN2 (Karras et al., 2020) training, for which we use audio augmentations. The SC prior is smoother and more accurate given its larger training set ($\sim$100,000 clips, 2,418 speakers post split).

5 RESULTS

In this section, we discuss the attack performance of the proposed GAN-based method compared to the GAN-free baseline, challenged under several ill-posed axes: 1) Perturbed gradient, 2) Compressed MFCC features, 3) Trained Networks, and 4) Complex architectures beyond CNNs.

5.1 GRADIENT PERTURBATION ANALYSIS

Table 1: Attack performance under increasing perturbations. Model Acc.: defended victim accuracy.

			GAN-free (baseline)					GAN-based (proposed)				
Dataset	Noise (z)	Model Acc.	Mel MSE	STOI	PESQ	ECAPA Cos.Sim	Speaker ID %	Mel MSE	STOI	PESQ	ECAPA Cos.Sim	Speaker ID %
AM	No Def.	98.03%	0.233	0.929	3.076	0.784	98.0%	**0.181**	0.895	2.312	0.713	98.0%
	0	98.24%	0.427	0.933	3.122	0.787	99.0%	**0.402**	0.893	2.315	0.718	95.0%
	10^{-5}	97.97%	0.477	0.925	2.975	0.751	98.0%	**0.426**	0.891	2.231	0.698	97.0%
	10^{-4}	98.16%	0.523	0.898	2.363	0.697	98.0%	**0.395**	0.894	2.257	**0.719**	98.0%
	3×10^{-4}	98.37%	0.761	0.835	1.606	0.491	97.0%	**0.428**	0.877	**2.036**	**0.653**	95.0%
	10^{-3}	98.27%	2.515	0.680	1.141	0.164	18.0%	**0.611**	**0.784**	**1.407**	**0.405**	**84.0%**
SC	No Def.	79.94%	1.584	0.623	1.572	0.414	75.0%	**0.227**	**0.726**	**2.299**	**0.710**	**98.8%**
	0	79.94%	1.549	0.624	1.564	0.407	74.0%	**0.232**	**0.727**	**2.311**	**0.720**	**99.5%**
	10^{-5}	79.87%	1.535	0.627	1.580	0.417	76.0%	**0.233**	**0.727**	**2.304**	**0.717**	**99.3%**
	10^{-4}	79.83%	1.551	0.624	1.558	0.409	74.5%	**0.233**	**0.722**	**2.234**	**0.703**	**98.5%**
	3×10^{-4}	79.52%	1.713	0.601	1.413	0.340	65.8%	**0.245**	**0.717**	**2.164**	**0.689**	**98.3%**
	10^{-3}	73.53%	3.168	0.510	1.153	0.197	29.3%	**0.387**	**0.663**	**1.740**	**0.545**	**96.5%**

In Table 1, the client clips its gradient to a fixed norm C (the median per-sample gradient norm) and adds Gaussian noise with a standard deviation zC, as in a single DP-SGD step (Abadi et al., 2016), where $z = 0$ denotes pure clipping. As C is data-derived and no privacy accounting is applied, we treat this as a controlled perturbation rather than certified differential privacy (App. A.6). Under increasing perturbation strength, the GAN-based attack degrades gracefully, while the GAN-free attack collapses. Additionally, the GAN-based method consistently scores the lowest Mel-MSE.

Most significantly, at a $z = 10^{-3}$, the GAN-free baseline shows a speaker ID % drop to 18.0% for the AM dataset and to 29.3% for the SC dataset. However, the proposed GAN-based method sustains its speaker ID leakage at 84.0% and 96.5% respectively, despite the strongest tested perturbation. In terms of fidelity, the proposed method is dominant for the SC dataset, with the quality prior trained on the diverse dataset. However, the AM prior is also recovering intelligible audio and is superior in its performance under the strongest perturbations. As a key insight, identity leakage decouples from reconstruction quality: at $z = 3 \times 10^{-4}$ on AM, even the GAN-free attack retains 97.0% speaker ID while its PESQ roughly halves, and at $z = 10^{-3}$ the GAN-based attack retains 84.0% despite a PESQ of 1.41. Degraded content or fidelity is therefore not evidence of speaker privacy. Also, we observe that the Mel Cos.Sim remains high, around 0.99 for both attacks (App. A.4), while GAN-free attack performance silently degrades for stronger perturbations.

5.2 COMPRESSED MFCC FEATURE ANALYSIS

In this section, we observe that GAN-based attacks are much more effective at recovering audio from compressed MFCC feature gradients, as seen in Table 2. Notably, when the features are more compressed with a lower number of MFCC coefficients n_{mfcc}, the GAN-based attack sustains speaker identity leakage, while the GAN-free method under-performs.

Table 2: Attack performance under compressed MFCC features (undefended).

		GAN-free (baseline)					GAN-based (proposed)				
Dataset	n_{mfcc}	MFCC MSE	STOI	PESQ	ECAPA Cos.Sim	Speaker ID %	MFCC MSE	STOI	PESQ	ECAPA Cos.Sim	Speaker ID %
AM	40	1.064	0.858	1.816	0.594	99.0%	**0.227**	0.884	1.702	**0.670**	**100.0%**
	20	2.563	0.769	1.386	0.339	74.0%	**0.479**	**0.837**	1.379	**0.514**	**93.0%**
	15	4.000	0.726	1.318	0.250	48.0%	**0.581**	**0.824**	1.373	**0.421**	**83.0%**
SC	40	2.627	0.711	1.976	0.603	96.5%	**0.272**	0.717	2.006	**0.711**	**99.8%**
	20	5.904	0.584	1.393	0.289	55.3%	**0.445**	**0.676**	1.488	**0.526**	**96.3%**
	15	6.288	0.559	1.325	0.219	34.0%	**0.539**	**0.656**	1.395	**0.414**	**84.3%**

5.3 TRAINED NETWORK ANALYSIS

Unlike randomly initialised networks, trained models emit suppressed gradients as learning progresses (Geiping et al., 2020); so samples move the model parameters less and yield smaller gradients; hence, we use a cosine-dominant L_{grad}. Table 3 shows the attack performance in later epochs.

Table 3: Attack performance on victims trained for 10, 20 and 30 epochs (undefended).

		GAN-free (baseline)					GAN-based (proposed)				
Dataset	Epoch	Mel MSE	STOI	PESQ	ECAPA Cos.Sim	Speaker ID %	Mel MSE	STOI	PESQ	ECAPA Cos.Sim	Speaker ID %
AM	10	24.415	0.637	1.221	0.240	35.0%	**1.078**	0.750	1.280	0.364	**70.0%**
	20	28.091	0.599	1.128	0.203	33.0%	**0.720**	0.717	1.247	0.310	**55.0%**
	30	24.712	0.646	1.165	0.244	42.0%	**0.590**	0.741	1.257	0.338	**68.0%**
SC	10	10.424	0.684	2.467	0.592	73.3%	**4.690**	0.658	1.840	0.544	**82.5%**
	20	3.672	0.756	2.525	0.726	94.3%	**0.817**	0.719	2.160	0.674	**96.8%**
	30	13.846	0.686	1.608	0.572	94.8%	**1.127**	0.700	1.929	0.601	**95.5%**

5.4 DIVERSE ARCHITECTURE ANALYSIS

Prior GIA work (Li et al., 2023; Zeng & Rudzicz, 2025; Dang et al., 2022) only evaluate one victim per setting: CNNs, LeNet style architectures and ASR models. In Table 4, we observe the attack performance under three further victim models. The GAN-based attack recovers speaker identity on LeNet where the GAN-free attack fails, but both attacks leak little identity on ResNet18 and AST. We attribute this to how each architecture aggregates its input before the classifier head. AudioCNN and LeNet use local max pooling retaining time-frequency positions. ResNet18 applies global average pooling, which discards absolute time–frequency position before the head; we hypothesise that global attention in AST behaves similarly. In both cases the gradient becomes more input invariant.

Table 4: Attack performance on untrained, undefended victims of different architectures.

		GAN-free (baseline)					GAN-based (proposed)				
Dataset	Model	Mel MSE	STOI	PESQ	ECAPA Cos.Sim	Speaker ID %	Mel MSE	STOI	PESQ	ECAPA Cos.Sim	Speaker ID %
AM	LeNet	14.116	0.439	1.051	0.094	2.0%	**0.403**	**0.771**	**1.313**	**0.456**	**91.0%**
	ResNet18	25.182	0.191	1.044	0.043	1.0%	**14.634**	0.344	1.055	0.066	3.0%
	AST	13.501	0.214	1.049	0.050	1.0%	**2.852**	**0.471**	**1.120**	**0.120**	**17.0%**
SC	LeNet	16.901	0.308	1.058	0.089	4.8%	**0.539**	**0.634**	**1.496**	**0.470**	**87.5%**
	ResNet18	28.553	0.189	1.054	0.067	4.3%	**16.128**	0.248	1.063	0.050	3.3%
	AST	12.274	0.174	1.066	0.088	6.0%	**5.345**	**0.236**	**1.141**	**0.136**	**15.0%**

6 ABLATIONS

The two main improvements of the proposed work include; the use of a GAN to generate realistic mel-spectrograms, and the use of BVG (Lee et al., 2023) in place of the conventional GL algorithm (Griffin & Lim, 1984). In the following sections, we discuss the improvement gained by the use of BVG, the upper-bound of the performance expected from the GAN latent space constraint. Furthermore, we explore the cross-dataset prior performance, to test feasibility of using general priors without strong assumptions on in-distribution data.

6.1 VOCODER IMPROVEMENT IMPACT

In this section, we discuss the improvement gained by switching to BVG vocoder. The GL algorithm relies on mathematical phase estimation, whereas BVG, trained on a large volume of audio data, is a good zero-shot vocoder. In Table 5, we observe that the attack performance is better with BVG across all metrics. The roundtrip measures waveforms converted into mel-spectrograms and back.

Table 5: Vocoder ablation. Roundtrip: ground-truth mel-spectrograms vocoded back to waveforms.

Dataset	Variant	Vocoder	STOI	PESQ	ECAPA Cos.Sim
AM	Roundtrip	GL	0.914	2.913	0.821
		BVG	**0.985**	**4.255**	**0.968**
	GAN-based	GL	0.853	2.010	0.636
		BVG	**0.895**	**2.312**	**0.713**
SC	Roundtrip	GL	0.786	3.415	0.905
		BVG	**0.872**	**4.148**	**0.966**
	GAN-based	GL	0.700	2.153	0.698
		BVG	**0.726**	**2.299**	**0.710**

6.2 GAN PRIOR UPPER-BOUND

In Table 6, we observe the upper-bound set on the expected attack performance due to the use of the GAN prior in the attack. The direct mel search bypasses the gradient matching objective and optimises the GAN latent to directly produce the ground truth mel-spectrogram. Note that, despite never having trained on victim speakers, the GAN manifold can represent unseen speaker identities accurately, resulting in the enhanced identity leakage in the attack.

Table 6: Empirical upper-bound on the ability of the GAN to produce the ground truth mel spectrograms versus the true attack performance along with cross-dataset prior performance

Dataset	Prior Dataset	Variant	Mel MSE	STOI	PESQ	ECAPA Cos.Sim	Speaker ID %
AM	AM	GAN-based Attack (proposed)	0.181	0.895	2.312	0.713	98.0%
		Direct mel search in GAN	0.028	0.945	2.968	0.900	100.0%
	SC	GAN-based Attack (proposed)	0.220	0.881	2.061	0.671	96.0%
		Direct mel search in GAN	0.050	0.927	2.620	0.830	100.0%
SC	SC	GAN-based Attack (proposed)	0.227	0.726	2.299	0.710	98.8%
		Direct mel search in GAN	0.087	0.764	2.874	0.812	100.0%

Cross-dataset prior attack performance: We also run an additional experiment by using the GAN prior trained on the SC dataset to conduct an attack on the AM dataset. Note that the SC dataset includes the same digit (0-9) classed audio clips as a subset, similar to the AM dataset. This version of the attack also yields comparable attack performance, as seen in Table 6, relaxing the need to have access to in-distribution victim data.

7 DISCUSSION & CONCLUSION

In this work, we show that a learnt audio prior can preserve speaker identity leakage when gradient inversion becomes information-limited. Under gradient perturbation and compressed MFCC features, the GAN-based attack retains speaker identity at levels where the GAN-free baseline loses it, even as fidelity degrades, while victims with global pooling or attention (ResNet18, AST) remain considerably more resistant. Poor content reconstruction alone is therefore not sufficient evidence of speaker privacy, and defences should be evaluated on identity leakage directly. Censor (Zhang et al., 2025a), which perturbs the update orthogonally to the true gradient, suppresses both content and identity for both attacks (App. A.6), showing that identity-concealing defences are attainable. Therefore, we highlight the need to study vulnerabilities in the audio domain further, in order to develop identity agnostic models and defences that not only suppress content but also conceal identity. As future work, extending the utility of GAN priors in audio GIA to long and variable length audio sequences and batched updates, similar to GAN-free audio work (Zeng & Rudzicz, 2025), remains an open direction. Furthermore, the potential of GAN-based methods to be extended to inference-time audio inversion (Zhang et al., 2025b) (App. A.1) needs to be explored as well.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods: the source code for the attack pipeline, model training, evaluation, and figure-plotting scripts was written with assistance from a large language model (Anthropic Claude, Opus). All such code was reviewed by an author and validated with correctness gates, assertions and sanity checks; the figures themselves were not produced by AI, only the plotting code that generates them. We also used generative AI to support the interpretation of results, specifically to cross-check and stress-test claims that were formulated by the authors.

We have not used generative AI tools for the core research ideas, the novelty, or the design of the methodology and experiments, including the selection of model components, all of which are author-conceived; AI assistance in this area was limited to surveying candidate options and prior literature. We did not use generative AI to generate synthetic data: no benchmark dataset was fabricated or augmented by an AI tool. The generative mel-spectrogram prior studied in this paper is a component of the attack method under investigation, not a tool used in preparing the submission. Formulating mathematical claims and assisting with proofs are not applicable to this work, which contains no formal theoretical results.

Additionally, we used generative AI tools for tasks with recommended disclosure. The core manuscript is author-written; large language models (Anthropic Claude, Opus and OpenAI ChatGPT, 5.x) were used to paraphrase and edit author-written text, to draft parts of the appendix and the disclosure statements from author-provided content, and to expand literature coverage. The literature review itself was conducted by the authors.

We have reviewed all AI-assisted work: AI-assisted code was verified and tested by an author along with correctness gates and assertions, AI-edited text was checked against the authors' intended meaning, and AI-supported claim checks were confirmed against our experimental results. We take responsibility for the final content of this work, including all text, claims, or artefacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies an offensive privacy attack, and we acknowledge its dual-use nature. Our intent is defensive: we characterise a leakage risk that persists under content-suppressing gradient perturbation in order to motivate stronger protection for audio federated learning, not to enable misuse. We recommend that practitioners combine state-of-the-art defences with more realistic federated learning configurations, such as multi-step local updates and large-batch aggregation, which increase the difficulty of inversion.

All experiments use two publicly available benchmark datasets, AudioMNIST (MIT licence) and Speech Commands v2 (CC-BY 4.0), under their respective licences. No clinical, private, or newly collected human-subject data is used, and no IRB approval was required. We operate under the standard honest-but-curious server threat model in federated learning. Our evaluation quantifies speaker-identity leakage only as embedding-level consistency between a reconstruction and its target; at no point do we attempt to determine, link, or reveal the real-world identity of any speaker. The societal relevance of this study lies in sensitive applications of audio federated learning, such as health-related speech analysis, where our findings show that suppressing content or fidelity does not, by itself, guarantee speaker privacy.

To preserve double-blind review, funding sources and affiliations are not disclosed here. We declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure this work is reproducible, with the supporting material provided across the main paper, appendix, and supplementary materials. The attack method and its two-stage pipeline are described in Section 3 and Algorithm 1, with the complete attack and ILO optimisation hyper-parameters in Tables 12 and 13. The mel-spectrogram extraction contract, fixed across all experiments, is given in Table 14, and the victim architectures in Tables 15 and 16. The datasets are public; their processing and the identity-disjoint speaker split are described in Sec-

tion 4.1, and the exact split files are provided as supplementary material. The recalibration of the gradient-perturbation grid for the single-sample threat model is explained in App. A.6. Random seeds are fixed and logged for all reported runs, and third-party component versions are pinned in the released code. Experiments are hardware-agnostic and run on commodity GPUs, with NVIDIA L4 and L40S recommended. We provide anonymized source code and trained GAN prior checkpoints for review [2].

ACKNOWLEDGMENTS

Placeholder for anonymous submission.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pp. 308–318, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 440–445, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1045. URL https://aclanthology.org/D17-1045/.

Mislav Balunović, Dimitar I. Dimitrov, Nikola Jovanović, and Martin Vechev. Lamp: extracting text from gradients with language model priors. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428, 2024. ISSN 0016-0032. doi: https://doi.org/10.1016/j.jfranklin.2023.11.038. URL https://www.sciencedirect.com/science/article/pii/S0016003223007536.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

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)*, pp. 4338–4342. IEEE, 2022.

Giannis Daras, Joseph Dean, Ajil Jalal, and Alexandros G. Dimakis. Intermediate layer optimization for inverse problems using deep generative models. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231925054.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, pp. 3830–3834, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[2] https://osf.io/pquwe/overview?view_only=a802c106f44b4b36ab33d7b2d4a9e941

EU. Regulation (EU) 2016/679 of the European Parliament and of the Council, 2016. URL https://data.europa.eu/eli/reg/2016/679/oj.

Hao Fang, Bin Chen, Xuan Wang, Zhi Wang, and Shu-Tao Xia. Gifd: A generative gradient inversion method with feature domain optimization. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4944–4953, 2023. doi: 10.1109/ICCV51070.2023.00458.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael 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*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Interspeech 2021*, pp. 571–575, 2021. doi: 10.21437/Interspeech.2021-698.

D. Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984. doi: 10.1109/TASSP.1984.1164317.

Samyak Gupta, Yangsibo Huang, Zexuan Zhong, Tianyu Gao, Kai Li, and Danqi Chen. Recovering private text in federated learning of language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Ali Hatamizadeh, Hongxu Yin, Holger Roth, Wenqi Li, Jan Kautz, Daguang Xu, and Pavlo Molchanov. Gradvit: Gradient inversion of vision transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10011–10020, 2022. doi: 10.1109/CVPR52688.2022.00978.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Jinwoo Jeon, Jaechang Kim, Kangwook Lee, Sewoong Oh, and Jungseul Ok. Gradient inversion with generative image prior. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, 2020.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Sanggil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. BigVGAN: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhuohang Li, Jiaxin Zhang, Luyang Liu, and Jian Liu. Auditing privacy defenses in federated learning via generative gradient leakage. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10122–10132, 2022. doi: 10.1109/CVPR52688.2022.00989.

Zhuohang Li, Jiaxin Zhang, and Jian Liu. Speech privacy leakage from shared gradients in distributed learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095443.

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*, 8(3):1798–1823, 2015. doi: 10.1137/14098435X. URL https://doi.org/10.1137/14098435X.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Eloi Moliner, Jaakko Lehtinen, and Vesa Välimäki. Solving audio inverse problems with a diffusion model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Pretom Roy Ovi and Aryya 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)*, pp. 4835–4839, 2024. doi: 10.1109/ICASSP48485.2024.10445809.

A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pp. 749–752 vol.2, 2001. doi: 10.1109/ICASSP.2001.941023.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. ISSN 0167-2789. doi: https://doi.org/10.1016/0167-2789(92)90242-F. URL https://www.sciencedirect.com/science/article/pii/016727899290242F.

Daniel Scheliga, Patrick Mäder, and Marco Seeland. Dropout is not all you need to prevent gradient leakage. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 9733–9741, 2023.

Jingwei Sun, Ang Li, Binghui Wang, Huanrui Yang, Hai Li, and Yiran Chen. Soteria: Provable defense against privacy leakage in federated learning from representation perspective. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315, 2021. doi: 10.1109/CVPR46437.2021.00919.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214–4217, 2010. doi: 10.1109/ICASSP.2010.5495701.

Soroosh Tayebi Arasteh, Cristian David Ríos-Urrego, Elmar Nöth, Andreas Maier, Seung Hee Yang, Jan Rusz, and Juan Rafael Orozco-Arroyave. Federated Learning for Secure Development of AI Models for Parkinson's Disease Detection Using Speech from Different Languages. In *Interspeech 2023*, pp. 5003–5007, 2023. doi: 10.21437/Interspeech.2023-2108.

Soroosh Tayebi Arasteh, Mahshad Lotfinia, Paula Andrea Perez-Toro, Tomas Arias-Vergara, Mahtab Ranji, Juan Rafael Orozco-Arroyave, Maria Schuster, Andreas Maier, and Seung Hee Yang. Differential privacy enables fair and accurate ai-based analysis of speech disorders while protecting patient data. *npj Artificial Intelligence*, 1(1):37, 2025.

US Congress. Health Insurance Portability and Accountability Act of 1996. Pub. L. 104-191, 110 Stat. 1936, 1996. Codified as amended at scattered sections of 18, 26, 29, and 42 U.S.C.

P. Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *ArXiv e-prints*, April 2018. URL https://arxiv.org/abs/1804.03209.

Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M. Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16332–16341, 2021. doi: 10.1109/CVPR46437.2021.01607.

Xijie Zeng and Frank Rudzicz. How to recover long audio sequences through gradient inversion attack with dynamic segment-based reconstruction. In *Proc. Interspeech 2025*, pp. 5118–5122, 2025.

Kaiyuan Zhang, Siyuan Cheng, Guangyu Shen, Bruno Ribeiro, Shengwei An, Pin-Yu Chen, Xiangyu Zhang, and Ninghui Li. Censor: Defense against gradient inversion via orthogonal subspace bayesian sampling. In *32nd Annual Network and Distributed System Security Symposium, NDSS 2025*, 2025a.

Xianglong Zhang, Haoming Luo, Mingda Han, Qihao Dong, Guoming Zhang, Yanni Yang, Qianli Li, and Pengfei Hu. Invercrs: Generative audio inversion attack in collaborative recognition systems. In *2025 IEEE 45th International Conference on Distributed Computing Systems (ICDCS)*, pp. 516–526, 2025b. doi: 10.1109/ICDCS63083.2025.00057.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *ArXiv*, abs/2001.02610, 2020. URL https://api.semanticscholar.org/CorpusID:210064455.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper-files/paper/2019/file/60a6c4002cc7b29142def8871531281a-Paper.pdf.

A Appendix

A.1 Extended related work

Table 7 summarises the closest audio inversion works. HFGM (Dang et al., 2022) reconstructs the MFCC input of an ASR model from its last-layer gradient. The method identifies speakers from the reconstructed features with an in-house identifier that reaches 42% top-1 even on original audio. Its identity leakage collapses under DP-SGD at noise multiplier 1×10^{-3} and under dropout. SPL (Li et al., 2023) formalises two-stage feature and waveform recovery on a keyword-spotting CNN and serves as our GAN-free baseline, re-implemented under our mel contract (80×86 versus SPL's 32×32 inputs) with BVG vocoder for both methods. Therefore, the absolute scores differ, but both settings show the GAN-free attack losing identity on MFCC inputs (Table 2). LAS (Zeng & Rudzicz, 2025) extends SPL to long and batched sequences through segment-wise reconstruction, which is orthogonal to the use of generative priors but it can be extended to handle long/batched audio as future work. The headline mel-spectrogram cosine similarity metric in LAS is uninformative in our setting (App. A.4). Both SPL and LAS report ECAPA-TDNN similarity without threshold calibration, unlike our FAR/EER-calibrated verification with a negative control (App. A.3). Ovi & Gangopadhyay (2024) also report audio gradient inversion, but we were unable to access the full text. Authors Tayebi Arasteh et al. (2025) show in a single-sample proof of concept that DP-SGD could suppress iDLG (Zhao et al., 2020) reconstruction of speech but without measuring identity. We leave dropout to future work, as dropout-aware inversion has been proposed for images (Scheliga et al., 2023) and could be combined with a generative prior. InverCRS (Zhang et al., 2025b) attacks the split-inference activations with a trained decoder and does not measure speaker identity. A pre-trained mel-spectrogram prior similar to ours could apply a manifold constraint on its decoder. CQT-Diff (Moliner et al., 2023) applies a diffusion prior to restore audio where the degradation is known. In contrast, our attack is conducted on a gradient perturbed by a defence unknown to the attacker (Sec. 3.1). The diffusion priors are therefore an alternative to our GAN prior. Finally, we inherit the W^+ and ILO (Daras et al., 2021) design of image-domain generative attack GIFD (Fang et al., 2023), and, as GradViT (Hatamizadeh et al., 2022), LAMP (Balunović et al., 2022), and FILM (Gupta et al., 2022) show that each architecture and modality needs its own adaptation, we contribute

the audio-specific components: an identity-disjoint mel-spectrogram prior, a differentiable MFCC path, and neural vocoding improvement.

Table 7: Closely related audio inversion work, grouped by the attacked signal. Only our attack combines a learnt prior, a neural vocoder and an evaluation on training gradients in audio FL. †Decoder trained on intercepted activations, not a prior over audio.

Attacked Signal	**Method**	**Input**	**Prior**	**Waveform synthesis**	**Defences evaluated**
Training Gradients (FL)	HFGM (Dang et al., 2022)	MFCC	None	None	DP-SGD, Dropout
	SPL (Li et al., 2023)	Mel, MFCC	None	Griffin-Lim	None
	LAS (Zeng & Rudzicz, 2025)	Mel	None	Griffin-Lim	None
	Proposed	Mel, MFCC	StyleGAN2	BigVGAN v2	Clipping with Gaussian noise, Censor
Split activations (inference time)	InverCRS (Zhang et al., 2025b)	Mel, MFCC	Trained decoder†	Phase estimation	Activation noise, Dropout
Degraded audio	CQT-Diff (Moliner et al., 2023)	Waveform	Diffusion model	Not needed	N/A

A.2 Qualitative reconstruction samples

Figures 2 and 3 show one held-out Speech Commands clip (for the word *five*) reconstructed by the GAN-free baseline and by the proposed GAN-based attack from the same shared gradient. Fig. 2 shows the mel-spectrograms as the FL setting becomes ill-posed under DP-SGD-style clipping and noise. Fig. 3 shows the MFCC diagrams as the victim gradient is produced by increasingly compressed MFCC features. As the gradient is perturbed or the feature is compressed, the GAN-free reconstruction loses the fine structure, while the GAN-based reconstruction retains the harmonic and formant structure in mel-spectrograms and the low-order cepstral bands in MFCCs. The difference panels (Δ) show the residual with reference to the ground truth. The audio clips used for the figures are attached in the supplementary material along with reconstructions of 3 other clips per dataset.

A.3 ECAPA metric analysis

We use the SpeechBrain ECAPA-TDNN checkpoint trained on VoxCeleb[3] and chose the threshold $\tau = 0.25$ based on its wide empirical use in standard libraries [4] [5]. However, we also conducted a post-hoc analysis on the threshold to identify key insights; 1) Per dataset False Acceptance Rate (FAR), Equal Error Rate (EER), 2) Negative control to test the effect of GAN fidelity improvement on ECAPA Cos.Sim. We also observe the genuine vs impostor distribution shifts for the LeNet style victim to visualise the effect of GAN-based attack on speaker identification improvement.

In Table 8, we observe the EER and FAR of the ECAPA verification threshold against the two ground truth datasets, which is attack agnostic. We consider the whole dataset for this evaluation with 60 speakers for the AM dataset and 2618 speakers for the SC dataset. The fixed threshold $\tau = 0.25$ accepts 11.6% of real *different-speaker pairs* on AudioMNIST and 2.7% on Speech Commands, signalling it is loose on the former and reasonable on the latter. The EER of ECAPA between two real one-second clips is 18.4 and 23.2% for SC and AM datasets, confirming the ECAPA embedding is intrinsically imprecise at this duration. However, the same threshold $\tau = 0.25$ is used to fairly evaluate both GAN-free and GAN-based methodologies.

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

[4] https://speechbrain.readthedocs.io/en/latest/API/speechbrain.inference.speaker.html#speechbrain.inference.speaker.SpeakerRecognition.verify_batch

[5] https://www.mathworks.com/help/audio/ref/speakerembeddings.html

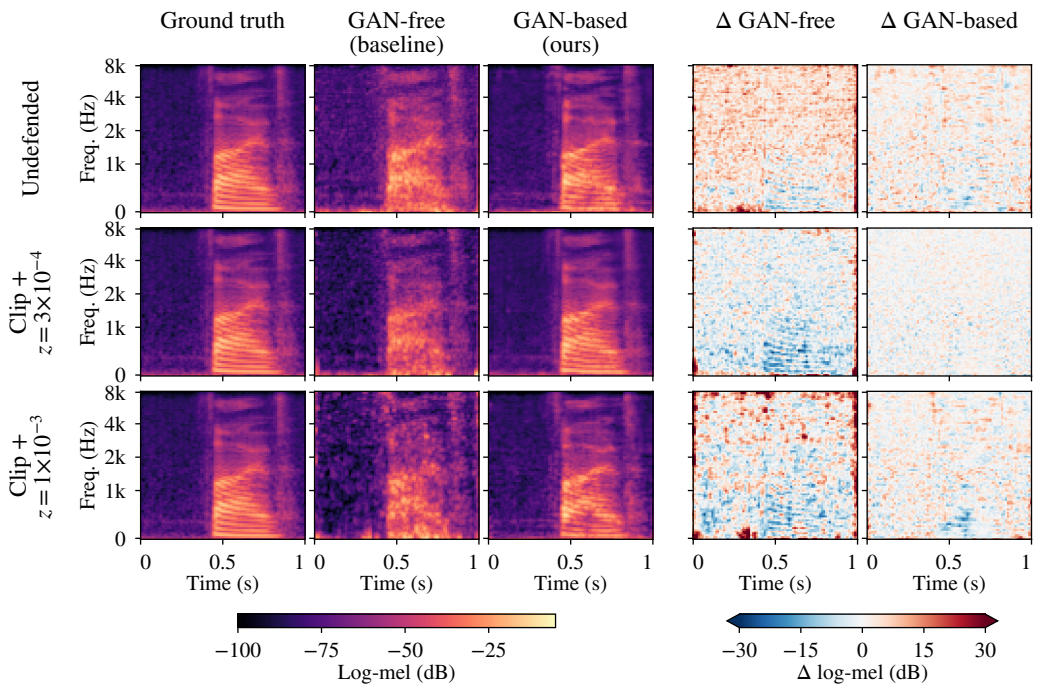


Figure 2: Log-mel spectrograms of a held-out Speech Commands clip (*five*) and its reconstructions, undefended (top) and under clipping with Gaussian noise multiplier $z = 3 \times 10^{-4}$ and $z = 10^{-3}$ (middle, bottom). Both attacks invert the same observed gradients and use the BVG vocoder. Δ: reconstruction − ground truth.

Table 8: ECAPA-TDNN EER and FAR on each dataset. Genuine pairs = same speaker, different clip; impostor pairs = different speaker; $\text{EER}_{\text{enroll}}$ is threshold-agnostic; FAR@0.25 is the fraction of different-speaker pairs the fixed threshold accepts. $\tau_{\text{FAR1\%}}$ is the threshold at which 1% of different-speaker pairs are accepted. Brackets: 95% confidence intervals.

Dataset	Genuine μ	Impostor μ	$\text{EER}_{\text{enroll}}$ (%)	FAR@0.25 (%)	$\tau_{\text{FAR1\%}}$ (ref.)
AM	0.288	0.126	23.2 [22.3, 24.0]	11.6 [11.0, 12.3]	0.409
SC	0.257	0.075	18.4 [17.9, 18.9]	2.7 [2.7, 2.8]	0.297

To rule out that the GAN's high-fidelity synthesis inflates ECAPA similarity, a negative control experiment is used. We embed 2000 unconditioned draws per prior and score them against averaged speaker enrolment embeddings, and observe the mean ECAPA Cos.Sim, FAR for the generated clips against real data at $\tau = 0.25$, and the $p95$ ECAPA Cos.Sim (Table 9). The negative control yields comparable 3.1% FAR and 9.1% for SC and AM datasets with reference to the ground truth FAR values of 2.7% and 11.6%. Also, the held-out and GAN's own train-set yield similar metrics. Therefore, the prior carries no usable identity even for the speakers it was trained on.

To evaluate if the prior shifts the genuine distribution under challenging settings, we observe the paired genuine and impostor distributions for the LeNet victim under both attacks in Figure 4. The GAN-free reconstructions largely overlap with the impostor distribution, while the GAN-based attack expands the gap between the genuine and impostor distributions. The Top-1 calculates the success rate of identification ranked by the cosine similarity to the average speaker ECAPA embedding. The effect is fidelity free as we leave out the ground truth reference clip for this evaluation. Notably, the speaker ID % (leak rate at $\tau = 0.25$) of the GAN-free method for the SC dataset is 4.8%, close to the 4.6% rate at which its reconstructions are accepted against other held-out speakers (impostor acceptance; cf. 2.7% for real clips, Table 8), with its Top-1 sitting at chance (0.5%),

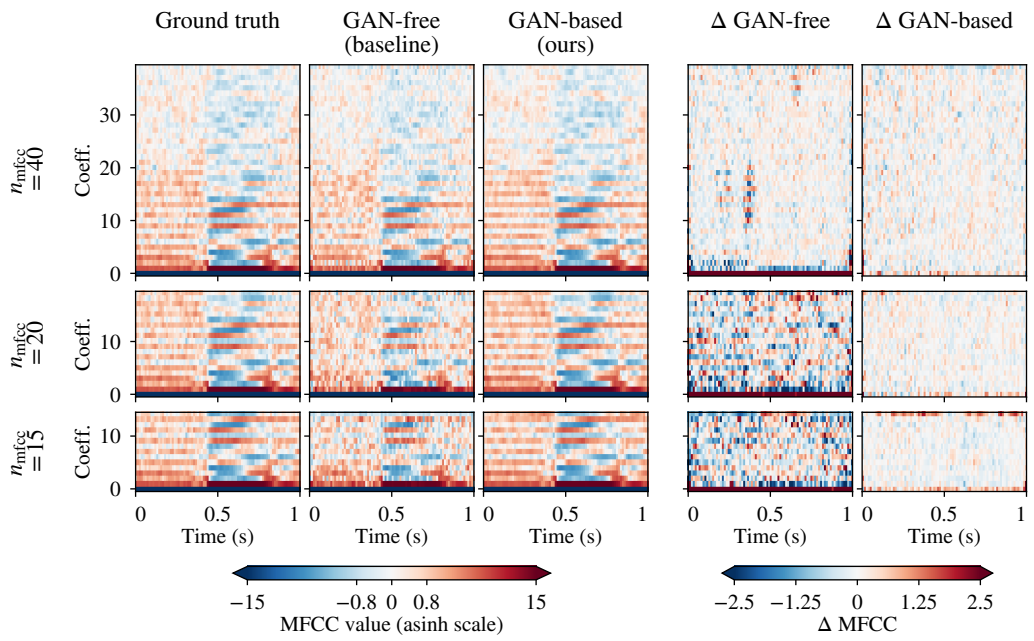


Figure 3: MFCCs of the same clip when the victim consumes the first $n_{\text{mfcc}} = 40$, 20 and 15 coefficients (top to bottom; undefended). Row height is proportional to n_{mfcc}, and c_0 is the bottom row of each panel. Δ: reconstruction − ground truth.

Table 9: Negative control: unconditioned draws per each prior (latent $\zeta \sim \mathcal{N}(0, I)$ without any optimisations), vocoded with BigVGAN and scored with ECAPA against the held-out speakers and against the speakers the prior was trained on. FAR@0.25 = fraction of pairs at or above 0.25. GT FAR@0.25 is FAR for ground truth different-speaker pairs on the same corpus (Table 8), for reference.

	Held out speakers			Prior trainset speakers			Per dataset (no prior)
Prior	Mean ECAPA Cos.Sim.	p95	FAR (%) @0.25	Mean ECAPA Cos.Sim.	p95	FAR (%) @0.25	GT dataset FAR@0.25
AM	0.1155	0.2892	9.1%	0.1171	0.2921	9.4%	11.6%
SC	0.0802	0.2270	3.1%	0.0818	0.2295	3.3%	2.7%

whereas the prior raises the leak rate to 88%, with Top-1 reaching 9.9% ($\approx 20\times$ chance). A similar effect can be observed for the AM dataset, with Top-1 rising to 51% ($\approx 5\times$ chance) for the prior compared to 14% for the GAN-free attack. The improvement is threshold-free, as AUROC is 0.98 for the GAN-based attack for both datasets versus 0.55–0.61 for the GAN-free baseline. The prior alone carries no identity (negative control); it enables the attack to extract the identity carried by the gradient.

A.4 Mel-spectrogram and MFCC cosine similarity

As seen in Table 10, the Mel Cos.Sim. metric does not vary significantly with different settings. Therefore, audio could remain unintelligible even with high cosine similarity. Although spectrograms are treated as images for the optimisation objective, they do not behave the same way when converted to audio. Unlike images that are relatively invariant to magnitude changes such as brightness, audio relies heavily on magnitude. The same phenomenon can be observed in Table 11 with MFCC Cos.Sim hovering around 0.99 for all scenarios.

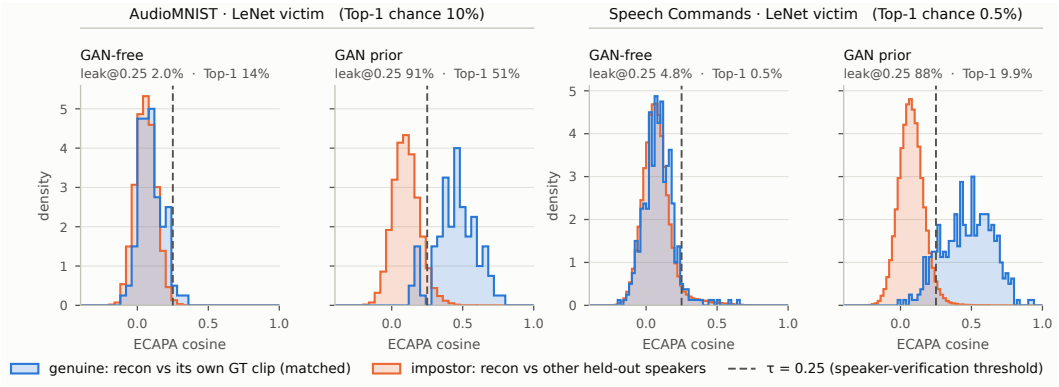


Figure 4: ECAPA Cos.Sim. of LeNet-victim reconstructions against their own speaker (genuine, blue) and other held-out speakers (impostor, red). The dashed line marks $\tau = 0.25$; panel titles give Speaker ID % and speaker embedding Top-1 accuracy.

Table 10: Mel Cos.Sim. holds around 0.99 while other metrics degrade substantially under heavy perturbation

		GAN-free (baseline)					GAN-based (proposed)				
Dataset	Noise (z)	Mel MSE	Mel Cos.Sim	STOI	PESQ	ECAPA Cos.Sim	Mel MSE	Mel Cos.Sim	STOI	PESQ	ECAPA Cos.Sim
AM	No Def.	0.233	**0.9995**	0.929	3.076	0.784	0.181	**0.9994**	0.895	2.312	0.713
	3.00E-04	0.761	**0.9985**	0.835	1.606	0.491	0.428	**0.9992**	0.877	2.036	0.653
	1.00E-03	2.515	**0.9922**	0.680	1.141	0.164	0.611	**0.9982**	0.784	1.407	0.405
SC	No Def.	1.584	0.9912	0.623	1.572	0.414	0.227	**0.9980**	0.726	2.299	0.710
	3.00E-04	1.713	0.9896	0.601	1.413	0.340	0.245	**0.9978**	0.717	2.164	0.689
	1.00E-03	3.168	0.9746	0.510	1.153	0.197	0.387	**0.9963**	0.663	1.740	0.545

A.5 Reproducibility

This section lists the settings needed to reproduce both attacks. Table 12 gives the optimisation hyper-parameters of Algorithm 1, Table 13 the ILO schedule of the GAN-based attack, and Table 14 the mel-spectrogram configuration shared by all components.

Attack hyper-parameters. Both attacks optimise $K = 4$ restarts and keep the one with the lowest final gradient-matching distance. The best-of-N probe applies only to the GAN-based attack, as it scores candidates drawn around w_{avg} in the generator latent space, which the GAN-free baseline does not have. The GAN-free baseline follows SPL (Li et al., 2023) with 8000 iterations at learning rate 0.01, while the GAN-based attack uses 3000 iterations, distributed over the ILO phases.

Table 11: MFCC Cos.Sim holds around 0.99 while other metrics degrade substantially under higher compression levels

		GAN-free (baseline)					GAN-based (proposed)				
Dataset	n_{mfcc}	MFCC MSE	MFCC Cos.Sim	STOI	PESQ	ECAPA Cos.Sim	MFCC MSE	MFCC Cos.Sim	STOI	PESQ	ECAPA Cos.Sim
AM	40	1.064	**0.9993**	0.858	1.816	0.594	0.227	**0.9997**	0.884	1.702	0.670
	20	2.563	**0.9993**	0.769	1.386	0.339	0.479	**0.9998**	0.837	1.379	0.514
	15	4.000	**0.9990**	0.726	1.318	0.250	0.581	**0.9998**	0.824	1.373	0.421
SC	40	2.627	**0.9966**	0.711	1.976	0.603	0.272	**0.9991**	0.717	2.006	0.711
	20	5.904	**0.9962**	0.584	1.393	0.289	0.445	**0.9995**	0.676	1.488	0.526
	15	6.288	**0.9962**	0.559	1.325	0.219	0.539	**0.9995**	0.656	1.395	0.414

The distance $\mathcal{D}$ combines the cosine distance, which matches the gradient direction (Geiping et al., 2020), with a down-weighted L_2 term. We keep the L_2 term because, unlike image intensities, mel-spectrogram magnitudes carry content and identity information (App. A.4), which a scale-invariant distance alone ignores. ATV smooths the mel-spectrogram more strongly along time than along frequency to preserve harmonic structure; for MFCCs it is applied along time only, as adjacent cepstral coefficients do not appear spatially smooth in real MFCCs.

Table 12: Attack optimisation hyper-parameters for Algorithm 1. Values are shared unless listed under a method-specific block; the GAN-free baseline uses the same K-restart protocol for parity.

Symbol	Description	Value
Shared		
K	Retained parallel restarts	4
–	Optimiser	Adam
$\mathcal{D}$	Grad-match distance	weighted L_2 + cosine
$(w_{L_2}, w_{\cos})$	L_2 / cosine weighting in $\mathcal{D}$	0.2 / 1.0
(λ_h, λ_v)	Mel ATV weights (time, freq), $\lambda_h > \lambda_v$	10^{-4} / 10^{-5}
(λ_h, λ_v)	MFCC ATV weights (time, freq)	10^{-4} / –
GAN-based only		
N	Probe pool (best-of-N init)	128
σ	Warm-start jitter std ($\epsilon \sim \mathcal{N}(0, \sigma^2)$)	0.5
$\mathcal{S}$	ILO layer schedule	Table 13
–	Optimisation iterations per restart	3000
GAN-free only		
η	Learning rate (single stage)	0.01
–	Optimisation iterations per restart	8000

ILO schedule. Phase 0 optimises the extended latent W^+ without a radius constraint, and phases 1–5 then successively unlock deeper intermediate generator layers s (Table 13); the per-phase iterations sum to the 3000-iteration budget. As optimisation descends to deeper, more expressive layers, we progressively tighten both the ℓ_1 trust-region radius r, which bounds the deviation of the active variable ϕ_s from its value when layer s is unlocked, and the ℓ_2 manifold proximity weight $\lambda_{\text{manifold}}$. Later phases therefore refine detail without drifting off the generator manifold (Fang et al., 2023).

Table 13: ILO phase schedule $\mathcal{S}$ for the GAN-based attack: per-phase iterations T, learning rate η, ℓ_1 trust-region radius r, and manifold (ℓ_2) proximity weight $\lambda_{\text{manifold}}$. Phase 0 optimises the latent W^+ (unconstrained radius); later phases optimise the intermediate activation at layer s.

Phase	Layer s	T	η	r (ℓ_1)	$\lambda_{\text{manifold}}$
0	W^+ (latent)	1500	0.05	–	0.00
1	1	400	0.02	256	0.03
2	2	400	0.01	128	0.05
3	3	300	0.01	64	0.10
4	4	200	0.01	32	0.15
5	5	200	0.01	16	0.20

Mel-spectrogram configuration. All components share the configuration in Table 14, which matches the input contract of the pretrained BigVGAN v2 vocoder, so reconstructions are vocoded without resampling or fine-tuning. The 80×86 log-mel is placed on the 128×128 GAN canvas under a fixed global affine normalisation per prior, which $\mathcal{C}$ inverts after cropping the real region (Sec. 3.3).

A.5.1 Victim model architectures

The architecture of the main victim model is described in Table 15. We use the same keyword spotting model as the one used in SPL (Li et al., 2023), modified to handle the mel configuration in this work. The additional architectures we use to evaluate our method are described in Table 16.

Table 14: Mel-spectrogram configuration parameters

Parameter	Value
Sample rate	22050 Hz
Clip length	1.0 s (22050 samples, fixed)
Window function	Hann
FFT size (n_{fft})	1024
Window length	1024
Hop length	256
Mel bins (n_{mels})	80
Minimum frequency (f_{min})	0 Hz
Maximum frequency (f_{max})	8000 Hz
Log magnitude floor	$\log(10^{-5})$
Real mel shape (freq $\times$ time)	80×86
GAN canvas (h $\times$ w)	128×128

Table 15: AudioCNN: base victim architecture used for attack evaluation. Output shapes shown for the 80×86 mel-spectrogram input; n_c is the number of classes (10 for AM, 35 for SC).

Layer	Kernel	Stride	Output
Input	–	–	$1 \times 80 \times 86$
Conv2d + ReLU	3×3	1	$32 \times 78 \times 84$
Conv2d + ReLU	3×3	1	$64 \times 76 \times 82$
MaxPool2d	2×2	2	$64 \times 38 \times 41$
Flatten	–	–	99,712
Linear (FC1) + ReLU	–	–	128
Linear (FC2)	–	–	n_c

A.5.2 GAN TRAINING

To verify that each mel prior has converged and to select the best checkpoint for inversion, we observe the GAN convergence. Since inception-based FID (Heusel et al., 2017) is not meaningful for mel-spectrograms, we use a classifier trained on the training data to generate embeddings that can be used to evaluate the true and GAN distributions for each coverage aspect. We tune the $R1$ regularization parameter to stabilise the trade-off between diversity and realism, sweeping $r1 \in \{2, 10, 20\}$ for 8,000 iterations. Convergence is considered under two main aspects: speaker coverage and content coverage over the vocabulary. The convergence curves for the SC and AM priors are shown in Fig. 5 and Fig. 6. The training and evaluation strictly exclude the held-out speaker clips to respect the id-disjoint split. For the SC dataset prior, we additionally measure coverage of the digit sub-manifold (*zero–nine*), depicted in Fig. 7, as it is used for the cross-dataset attack on the AM dataset. The AM dataset prior curves (Fig. 6) show noticeable fluctuations, whereas the SC prior curves (Fig. 5) are considerably smoother. This difference is a direct consequence of dataset scale:

Table 16: Additional victim architectures (Sec. 5).

Model	Description
LeNet	A conventional two-block convolutional network (channels $1 \to 6 \to 16$, 5×5 kernels), each block with ReLU and 2×2 max-pool, then two FC layers. The conv stack flattens directly into the FC.
ResNet18	18-layer deep residual network (He et al., 2016), first convolution adapted to single channel. Uses a global `AdaptiveAvgPool2d` before the head, which smears absolute frequency position and makes gradient inversion noticeably harder (Sec. 5).
AST	Audio Spectrogram Transformer (Gong et al., 2021): a ViT (Dosovitskiy et al., 2021) over non-overlapping 16×16 mel patches, learnable positional encoding, and a stack of self-attention blocks (parameters: token dim 512, depth 8, 8 heads).

AudioMNIST retains only ∼25,000 clips and 50 speakers after the held-out split (of 30,000 clips, 60 speakers), against ∼100,000 clips and 2,418 speakers (of 2,618) for Speech Commands.

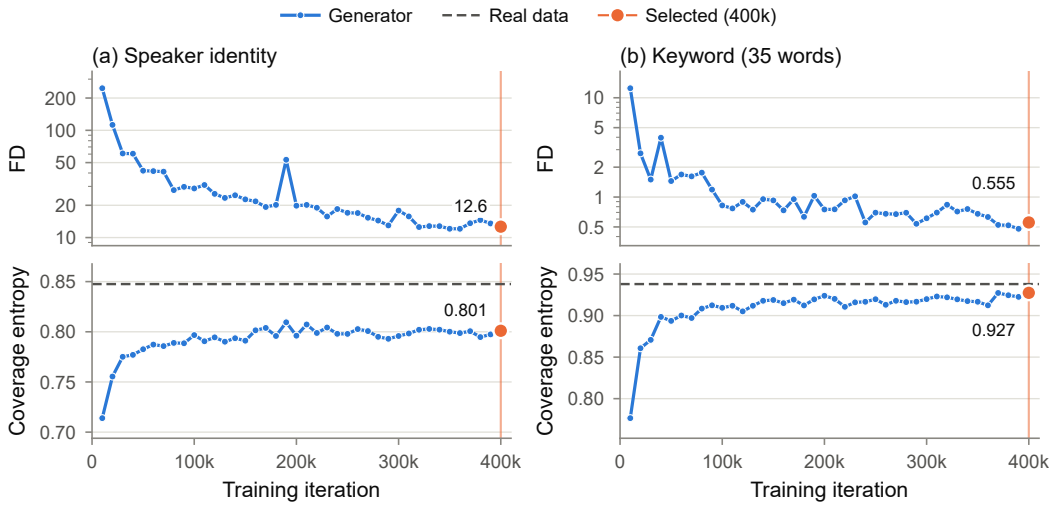


Figure 5: Convergence of the Speech Commands prior, scored with a classifier trained on real training-speaker mel-spectrograms for (a) speaker identity (2,418 speakers) and (b) keyword (35 words). Top: Fréchet distance (FD, log scale; lower is better) between 2,000 generator samples and the real training set. Bottom: normalised coverage entropy of the predicted classes (higher is better); the dashed line is the real-data value. Orange marks the selected checkpoint (400k).

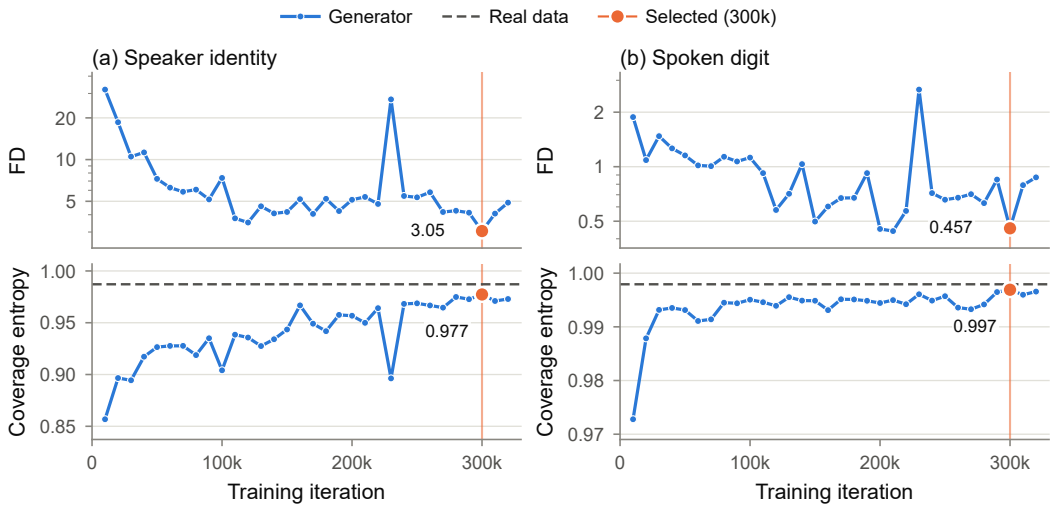


Figure 6: Convergence of the AudioMNIST prior, scored with a classifier trained on real training-speaker mel-spectrograms for (a) speaker identity (50 speakers) and (b) spoken digit. Top: Fréchet distance (FD, log scale; lower is better) between 2,000 generator samples and the real training set. Bottom: normalised coverage entropy of the predicted classes (higher is better); the dashed line is the real-data value. Orange marks the selected checkpoint (300k).

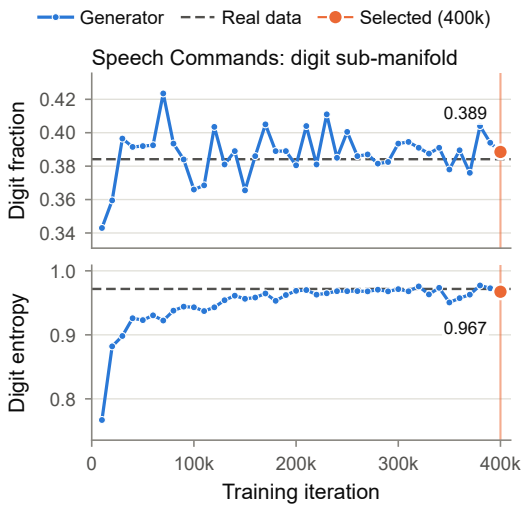


Figure 7: Digit sub-manifold coverage of the Speech Commands prior, which enables the cross-dataset attack on AudioMNIST. Top: fraction of 2,000 samples that the 35-way keyword classifier assigns to a digit word (*zero–nine*). Bottom: normalised entropy of the predicted digits (higher is more uniform). Dashed lines: real training data; orange: selected checkpoint (400k).

A.6 Defence setting

Clipping and Gaussian perturbation. Following the DP-SGD defence (Abadi et al., 2016), the client clips the shared gradient $g \in \mathbb{R}^D$ by its global L_2 norm over all parameters and adds Gaussian noise (Eq. 7).

$$g_{\text{obs}} = \mathcal{P}(g) = g \cdot \min\Big(1, \tfrac{C}{\|g\|_2}\Big) + \mathcal{N}\big(0,\, z^2 C^2 \mathbf{I}_D\big) \tag{7}$$

We set $C = 254$, the median $\|g\|_2$ of the untrained AudioCNN over a pilot set of targets, so that roughly half of the gradients are clipped. z is the noise multiplier, and $z = 0$ denotes pure clipping. The Model Acc. column of Table 1 reports the validation accuracy of separate victims trained with the same mechanism applied to each mini-batch gradient (batch size 64, C automatically set to the median batch-gradient norm before training, 50 epochs of SGD). No privacy accountant is used. Without batch averaging, the noise-to-signal ratio of a single clipped gradient is at least $z\sqrt{D}$. For AudioCNN ($D \approx 1.28 \times 10^7$), $z = 1$ would exceed the signal by over three orders of magnitude, so our grid instead spans ratios from 0.04 to 3.6 (Table 17).

Table 17: Noise multiplier z and the noise-to-signal ratio $z\sqrt{D}$ for the AudioCNN victim ($D \approx 1.28 \times 10^7$ for both datasets).

z	0	10^{-5}	10^{-4}	3×10^{-4}	10^{-3}
$z\sqrt{D}$	0	0.04	0.36	1.07	3.58

Censor. We evaluate Censor (Zhang et al., 2025a) with 10 candidate updates drawn orthogonally to the true gradient. Each candidate is rescaled to the per-layer norms of the true gradient and selected greedily by probe loss. Censor defeats both attacks (Table 18): Mel-MSE rises by over an order of magnitude, and ECAPA Cos.Sim falls to 0.06–0.07, at or below the impostor means of the ground-truth data (Table 8). Censor therefore conceals identity as well as content.

Future work. In exploratory runs, Soteria (Sun et al., 2021) and top-k sparsification (Aji & Heafield, 2017) perturbed the untrained AudioCNN gradient too little to separate the two attacks. Evaluating these defences at stronger settings, and on victims where they are effective, is left to future work.

Table 18: Attack performance under Censor (Zhang et al., 2025a) on the untrained AudioCNN victim.

	GAN-free (baseline)					GAN-based (proposed)				
Dataset	Mel MSE	STOI	PESQ	ECAPA Cos.Sim	Speaker ID %	Mel MSE	STOI	PESQ	ECAPA Cos.Sim	Speaker ID %
AM	37.079	0.307	1.053	0.058	0.0%	48.069	0.501	1.064	0.065	0.0%
SC	51.540	0.327	1.074	0.072	5.3%	57.872	0.367	1.101	0.062	5.5%