GAI:GENERATIVE MODEL INVERSION ATTACKS ON AUDIO SIGNALS

Devin Y. De Silva
Department of Computer Science
Rensslear Polytechnic Institute
Troy, NY 12180, USA

ABSTRACT

Deep neural networks used in Speaker Recognition Systems (SRSs) encode sensitive biometric information, raising significant privacy concerns regarding the potential leakage of training data. While Model Inversion (MI) attacks have been successfully demonstrated to reconstruct face images with high fidelity, their application to the audio domain has been limited. Existing audio MI techniques, which typically perform optimization directly on raw waveforms, often struggle to produce acoustically realistic or semantically intelligible speech. In this paper, we propose Generative Audio Inversion (GAI), a novel framework that adapts Generative Model Inversion (GMI) to the acoustic domain to reconstruct high-fidelity speaker representations from a target SincNet model. Our approach leverages a Generative Adversarial Network (GAN) trained on public auxiliary data to serve as a distributional prior, constraining the inversion search to the manifold of realistic mel-spectrograms. By optimizing in the latent space of the GAN and utilizing a neural vocoder for waveform synthesis, we perform evaluation on the proposed methods on biometric information recovered and distribution shift effects on the methodology. The code is available at `https://github.com/DevinDeSilva/AudioGenMI/tree/master`

1 INTRODUCTION

The inherent biometric nature of speech data raises significant privacy concerns for machine learning (ML) models used in speaker recognition. Speech parameters including accent, rhythm, and acoustic properties are known to carry sensitive personal informationKröger et al.. Consequently, ensuring that ML models for speaker recognition do not inadvertently leak information about their training data is a critical requirement.

However, recent studies have demonstrated that ML models are broadly susceptible to various privacy-compromising attacks Nasr et al.; Perez & Ribeiro; Jagielski et al. (a;b). A particularly relevant threat is the Model Inversion (MI) attack Zhang et al., which enables an adversary to reconstruct abstract representations of individual classes from the target model's training data. Given that speaker recognition systems typically treat each speaker as a distinct class, MI attacks pose a severe risk of privacy breach by potentially allowing the malicious retrieval of personal information about the speakers Pizzi et al. (2022). Despite the high risk, the practical feasibility of performing MI attacks on speaker recognition systems and audio data has is sparsely investigated, leaving the privacy vulnerability of speaker information in this context underexplored.

This challenge was significantly overcome in the image domain by the introduction of **Generative Model Inversion (GMI) attacks** Zhang et al.; Wang et al.. This approach departs from direct pixel-wise optimization by leveraging **Generative Adversarial Networks (GANs)** to learn an informative distributional prior from public, non-private datasets. By constraining the inversion search to the realistic data manifold learned by the GAN, GMI successfully regularized the optimization problem, enabling the hi gh-fidelity reconstruction of face images from state-of-the-art face recognition models Yin et al. (2023).

Recently, the model inversion threat was formally extended to the audio domain by Pizzi et al. Pizzi et al. (2022). Their work demonstrated that MI attacks could effectively extract audio samples and

d-vectors (intermediate voice feature representations) from a SincNet-based speaker recognition system Ravanelli & Bengio (2019). Their proposed **Sliding Model Inversion (SMI)** attack, which leverages the sequential properties of audio data, achieved up to 90% accuracy (claimed) in fooling the target classifier. While groundbreaking, the quality of the resulting audio reconstructions often lacked the fidelity required to fully fool a human listener, suggesting an avenue for improvement in utilizing the generative methodology as in the image space.

In this paper, we bridge the gap between the high-fidelity generative inversion demonstrated in the image domain and the emerging threat of audio model inversion. We propose a novel generative framework for MI attacks against speaker recognition systems, focusing on the reconstruction of high-quality, recognizable speech.

Our Contributions

Our work is centered on designing a robust, high-fidelity Model Inversion (MI) attack by adapting sophisticated generative techniques to the unique properties and challenges of speaker recognition systems. We present the following key contributions:

1. **Generative Model Inversion in the Audio Domain:** We successfully introduce and adapt the concept of Generative Model Inversion (GMI) to the audio domain.
2. **Generative Audio Inversion (GAI) Framework:** We propose a novel Model Inversion framework that utilizes a **Generative Adversarial Network (GAN)** trained on public, non-private data to synthesize realistic **mel-spectrograms**. This trained GAN serves as a powerful **generative prior**, which is instrumental in regularizing the inversion optimization process.
3. **Qualitative and High-Fidelity Assessment:** We qualitatively and quantitatively demonstrate the effectiveness of Generative Audio Inversion.

The remainder of this paper is organized as follows. Section 2 provides essential background on existing Model Inversion attacks and the architecture of the SincNet-based speaker recognition system. Section 3 details the design and optimization objectives of our Generative Audio Inversion framework. Section 4 describes the experimental setup, datasets, and the quantitative and qualitative evaluation metrics used. Section 5 presents the challenges and adaptation to there challenges.

2 Related Work

Our work focuses on evaluating the privacy risks of speaker recognition systems by introducing a novel Generative Model Inversion (GMI) attack tailored for acoustic data. This research draws upon foundational concepts from general machine learning privacy attacks, model inversion techniques in both the image and audio domains, and deep generative models.

2.1 Privacy in Speech and Audio

The deployment of Deep Neural Networks (DNNs) for **Speaker Recognition (SR)** has introduced significant privacy concerns Kröger et al.. Speech and voice signals inherently carry a rich array of sensitive biometric and paralinguistic information about the speaker, such as age, gender, personality traits, and physical health. It is crucial to distinguish between content privacy (protecting the semantic meaning of speech) and voice privacy (protecting the identity of the speaker). Recent works by Williams et al. have formalized this distinction, highlighting that standard anonymization often neglects the privacy risks embedded in the linguistic content itself Williams et al. (2022). While techniques such as VQ-VAE-based speech re-synthesis have been proposed to mask sensitive content words while preserving intelligibility Williams et al. (2024), our work addresses the complementary threat: the extraction of the speaker's biometric identity from the model itself.

2.2 Membership Inference Attacks (MIA)

The most recognized standard for assessing the privacy risks of machine learning models is the Membership Inference Attack (MIA) Shokri et al.. An MIA determines whether a specific data record was included in a model's private training dataset. The feasibility of these attacks is fundamentally tied to the tendency of DNNs to overfit on their training data Hu et al. (2022). In the audio domain, MIA has recently evolved beyond simple error-based metrics. For instance, Teixeira et al. demonstrated that for Automatic Speech Recognition (ASR) models, leveraging **loss-based features** combined with Gaussian and adversarial perturbations significantly improves the detection of training samples compared to black-box transcription errors Teixeira et al. (2024). While SLMIA-SR pioneered MIA for Speaker Recognition Systems (SRSs) at the speaker level Chen et al. (2024), our work targets the more aggressive **Model Inversion**, which aims to reconstruct the data rather than simply detect its presence.

2.3 Model Inversion Attacks in the Audio Domain

A more aggressive class of privacy threats is the Model Inversion (MI) attack, which aims to physically reconstruct or retrieve sensitive features of the training data corresponding to a target class Fredrikson et al. (2015b); Pizzi et al. (2022). Initially demonstrated on simpler models like linear regression Wu et al. (2016), traditional MI attacks struggled against deep neural networks due to the high-dimensionality and ill-posed nature of the optimization problem, often leading to unrealistic and blurry reconstructions Fredrikson et al. (2015a). The first work to successfully extend MI attacks to audio data targeted a state-of-the-art SincNet-based speaker recognition system Ravanelli & Bengio (2019). Pizzi et al. Pizzi et al. (2022) demonstrated the reconstruction of both **raw audio samples** and intermediate d-vectors (voice feature representations). They introduced Sliding Model Inversion (SMI), which iteratively inverts overlapping chunks of audio to exploit its sequential properties, significantly boosting the attack's success rate. Crucially, this prior work showed that the inverted audio could be used to generate initial deepfake samples for voice impersonation attacks, validating the security risk posed by MI.

2.4 Generative Model Inversion

To overcome the limitations of directly inverting high-dimensional data like images, the concept of Generative Model Inversion (GMI) was proposed by several approaches Zhang et al.; Wang et al.. GMI dramatically enhanced the fidelity of reconstructions, especially for complex data like face images. The core idea is to leverage a Generative Adversarial Network (GAN) trained on a non-private, auxiliary public dataset to establish an informative distributional prior Zhang et al.. This prior effectively regularizes the complex inversion problem by constraining the search space to the manifold of realistic, recognizable data. The initial success of MI on raw audio Pizzi et al. (2022) demonstrated feasibility, yet the complexity reduction approaches tested in the image space have not been tested in the audio space before. This has potential as frequency domain transformations have improved speech synthesis Shen et al. (2018). Our work is motivated by the critical necessity to integrate these advancements. We propose the first Generative Audio Model Inversion attack by adapting the generative prior concept Zhang et al. to the acoustic domain.

3 Methodology

In this section, we detail our proposed **Generative Audio Inversion (GAI)** framework. Unlike previous approaches that perform optimization directly on raw audio waveforms Pizzi et al. (2022), our method targets the inversion of abstract spectral representations (mel-spectrograms) generated by a pre-trained Generative Adversarial Network (GAN). This approach leverages the distributional prior of natural speech to regularize the inversion process, ensuring high-fidelity reconstructions.

3.1 Threat Model and Problem Definition

We assume a **white-box** attack scenario where the adversary has access to the target speaker recognition model, T, including its architecture and parameters (e.g., gradients). The target model T is a SincNet based system Ravanelli & Bengio (2019) that maps raw audio waveforms x to a speaker identity y.

The adversary's goal is to recover a recognizable audio sample x^* corresponding to a specific target speaker identity y_t, such that $T(x^*) = y_t$. We assume the adversary does not have access to the

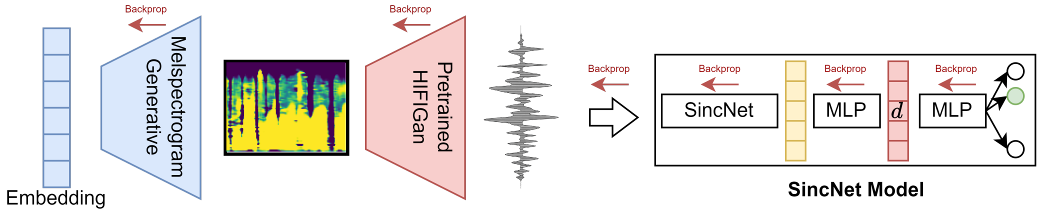


Figure 1: The **Generative Audio Inversion (GAI)** attack pipeline. During the inversion, the parameters of the Generator and HIFIGan are frozen to serve as a differentiable surrogate. The optimization process (indicated by red arrows) backpropagates the identity loss from the target SincNet model through the entire pipeline to iteratively update the latent embedding z, recovering the target speaker's biometric features.

private training data $X_{private}$ but has access to a generic, public auxiliary audio dataset X_{public} which is used to learn the generic features of human speech. We denote the mel-spectrogram of a audio sample x as $F_{mel}(x)$ where F_{mel} represents the mel-spectrogram conversion function

3.2 Generative Audio Prior Learning

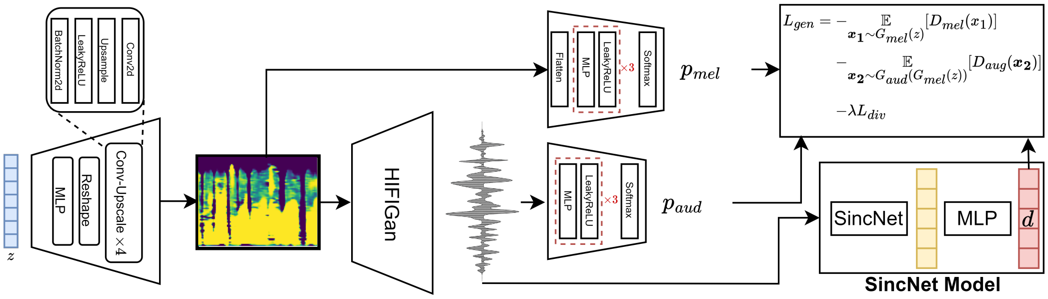


Figure 2: Schematic of the **Generative Prior Learning** phase. The framework maps a latent vector z to a Mel-spectrogram via a convolutional generator (G_{mel}) and subsequently to a waveform using a pre-trained HIFIGan. The generation process is regularized by a dual-discriminator setup (D_{mel} and D_{aud}) and a feature-matching diversity loss (L_{div}) extracted from the SincNet encoder, ensuring the synthesized audio aligns with the distribution of natural speech.

To constrain the optimization space to realistic audio, we first distill generic speech knowledge from the public dataset into a deep generative model. We employ a Generative Adversarial Network (GAN) as outlined in Gulrajani et al. (2017) to model the distribution of mel-spectrograms.

As illustrated in the figure above, our generative framework consists of two generators G_{mel}, G_{aud} and a pair of discriminators (D_{aud}, D_{mel}).

- **Mel Generator** (G_{mel})**:** The generator takes a low-dimensional latent vector z sampled from a normal distribution $\mathcal{N}(0, I)$ and outputs a synthetic mel-spectrogram $M = G_{mel}(z)$.
- **Audio Generator** (G_{aud})**:** The generator takes the generated mel-spectrogram M generated by the G_{mel} generator and outputs a synthetic audios $x^* = G_{aud}(M)$, During our implementation we utilized a HIFIGanKong et al. (2020) which is pretrained on LibriSpeech dataset Panayotov et al. (2015).
- **Discriminators** (D_{aud}, D_{mel})**:** To ensure the generated spectrograms and the audio generated conditioned on this mel-spectrogram is properly generated we utilize two discrimina-

tor. We utilize a image discriminator D_{mel} to assess the spectrogram and a audio discriminator D_{aud} to focus on generated audio signal (as shown in the Fig. 2).

The Generative models is trained on the auxiliary dataset to minimize the Wasserstein distance with gradient penalty Gulrajani et al. (2017), ensuring that for any random z, $G_{mel}(z)$ represents a realistic speech spectrogram and $G_{aud}(G_{mel}(z))$ represents the realistic audio signal. The Generators are trained using the loss function stated in 2 where L_{div} is defined in by Equation 1, F represents the feature vectors of the speaker recognition model (T) as proposed by Zhang et al.. While the discriminator pair is trained under the loss function mentioned in 2 where $\mathbb{P}_{\hat{\boldsymbol{x}}}$ denotes random samples generated by sampling uniformly along straight lines between pairs of points sampled from the data distribution $G_{aud}(G_{mel}(z))$ and X_{public}, D represents either D_{mel}, D_{aug} and $\mathbb{P}_g$ represent $G_{mel}(z), G_{aud}(G_{mel}(z))$ respectively.

$$L_{div} = \mathop{\mathbb{E}}_{\boldsymbol{z_1},\boldsymbol{z_2}\sim\mathcal{N}(0,I)}\left[\frac{\|F(G_{aud}(G_{mel}(\mathbf{z}_1)) - F(G_{aud}(G_{mel}(\mathbf{z}_2))\|}{\|\mathbf{z}_1 - \mathbf{z}_2\|}\right] \tag{1}$$

$$L_{dis} = \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_g}[D(\tilde{\boldsymbol{x}})] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim X_{public}}[D(\boldsymbol{x})] + \lambda \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}}\sim\mathbb{P}_{\hat{\boldsymbol{x}}}}\left[(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1)^2\right] \tag{2}$$

3.3 Generative Audio Inversion Attack

Once the generators G_{mel}, G_{aud} are trained, we freeze its parameters and perform the model inversion attack. The attack is formulated as an optimization problem in the latent space of the generator, rather than the pixel (or time-sample) space of the audio. The algorithm is provided in the appendix 1

The inversion pipeline, depicted above in Fig. 1, consists of three main stages:

1. **Latent Projection:** We initialize a random latent vector z. This vector is passed through the frozen Generator G_{mel} and G_{aud} to produce a candidate audio $x^* = G_{aud}(G_{mel}(z))$.
2. **Target Feedback and Optimization:** The synthesized audio x^* is fed into the target SincNet model. We compute the loss with respect to the target speaker class y_t and back propagate the gradients through the entire differentiable pipeline (SincNet → Vocoder → Generator) to update the latent vector z.

Following the principles of Generative Model Inversion Zhang et al., we search for the optimal latent vector z^* by minimizing the identity loss associated with the target speaker. The optimization problem is formulated as :

$$z^* = \text{argmin}_z \mathcal{L}_{id}(z, y_t) + \lambda \mathcal{L}_{prior}(z) \tag{3}$$

- **Identity Loss ($\mathcal{L}_{id}$):** This term guides the optimization to produce audio features that are classified as the target speaker y_t with high confidence. We employ the cross-entropy loss (or negative log-likelihood) between the target model's prediction $T(G_{aud}(G_{mel}(z)))$ and the target label y_t to maximize the likelihood of the target class.
- **Prior Loss ($\mathcal{L}_{prior}$)** This term guides the optimization to produce audio features and mel features that are realistic where $\mathcal{L}_{prior}(z) = -D_{mel}(G_{mel}(z)) - D_{aug}(G_{aud}(G_{mel}(z)))$.

By optimizing z directly within the latent space of the frozen generator, we leverage the learned distributional prior of the GAN to recover a mel-spectrogram that is acoustically structured and capable of fooling the SincNet classifier.

4 Experimental Design & Initial Results

4.1 Setup

We utilize VCTKGarofolo (1993a) as the main dataset to run the the experiments. The dataset contains WAV files of 151 speakers. Further we also have the information of the gender, Age, Accent and region of the speakers thus we are capable of identifying whether there is a leak in biometric information when we attack a model trained on this dataset. As a preprocessing step we create a Age group for speakers by categorizing the speakers with age less than 20, age between 20-28 and speakers aged above 28.

The sincNet model (T) was trained on 35 selected speakers from the origin 151 speakers available in the Garofolo (1993a) which will be our $X_{private}$. The speakers were selected such that the we select the speaker from each gender (of 2), each age group (of 3) , each Accent (of 11) which has the maximum duration of audio. As our public data X_{public} we utilize the rest of the speaker audios to train the generator ensuring that the speakers of the public and private set are disjoint.

We evaluate our methodology against the standard model inversion (STD) and sliding model inversion algorithms (SLIDING) stipulated in Pizzi et al. (2022). To create the inverted samples we initialize vectors in three different settings. Full of Zeros , laplacian distribution with $\mu = 0, b = 0.07$ and uniform noise within $[-1, 1]$

4.2 Comparative Results: Biometric Information Extraction

To evaluate how much the inverted samples captured biometric information, we create 3 inverted samples per speaker for per activation type for each method of evaluation. The inverted samples are evaluated by using classifiers that were trained to classify the given attacked speaker sample on the sensitive attribute (gender , age and accent). We consider that if the sample of a speaker is properly identified on the sensitive attribute, that sensitive attribute is leaked. Table 1.

Table 1: Gender based Model Inversion accuracy.

ID	Method	Init Type	Gender (%)	Age (%)	Accent (%)	Time (s)
1	STD	Zero	60.00	11.42	11.42	1.058
2	SLIDING	Zero	54.28	11.43	**12.38**	3.805
3	**GAI**	Zero	**66.67**	**21.90**	7.62	10.672
4	STD	Laplacian	57.14	20.95	7.62	1.277
5	SLIDING	Laplacian	52.38	16.19	**11.43**	2.910
6	**GAI**	Laplacian	**60.95**	**21.90**	6.67	10.779
7	STD	Uniform	52.38	20.95	8.57	1.161
8	SLIDING	Uniform	**61.90**	21.90	8.57	2.715
9	**GAI**	Uniform	47.62	**30.48**	**12.38**	13.510

As shown in the Table 1 **GAI** achieves higher accuracy on several attributes proving that a generative prior improves the ability to attack a speaker recognition model though this comes at the cost of increased attack time as it takes about 10x the standard attack.

4.3 Dataset Shift Analysis: Voxceleb1 dataset

To evaluate the effect of data shift on the public data compared to the private data we train our generator pair (G_{mel}, G_{aud}) on VoxCeleb1 Nagrani et al. (2019) and do the Generative model inversion attack on the speaker recognition model trained with VCTK Garofolo (1993a) dataset

Table 2: Gender based Model Inversion accuracy.

ID	Method	Init Type	Gender (%)	Age (%)	Accent (%)
1	VCTK $\rightarrow$ VCTK	Zero	66.67	21.90	7.62
2	VoxCeleb1 $\rightarrow$ VCTK	Zero	40.00	11.43	11.43
3	VCTK $\rightarrow$ VCTK	Laplacian	60.95	21.90	6.67
4	VoxCeleb1 $\rightarrow$ VCTK	Laplacian	49.52	16.19	6.67
5	VCTK $\rightarrow$ VCTK	Uniform	47.62	30.48	12.38
6	VoxCeleb1 $\rightarrow$ VCTK	Uniform	49.52	11.43	6.67

As the Table 2 suggest having a data-shift most of the time reduces the information that can be extracted. Further comparing the accuracy values of the STD and SLIDING from Table 1 we notice that the accuracy is less than the baselines thus suggesting that the attack heavily depend on the closeness of the public and private data distributions.

4.4 Combining attacks: **GAI** + STD

We evaluate the effectiveness of combining attacks by using the produced inverted audio sample from our methodology **GAI** as the input to the STD algorithm Pizzi et al. (2022). We notice that this improves the attack accuracies. The Table 3 shows the accuracies as well as the analysis of whether this improved or worsened the highest accuracy. ↑ indicate whether the accuracy was higher compared to the best accuracy in Table 1 or − if it's equal to the highest accuracy while ↓ indicate that the accuracy is lower than the highest accuracy in Table 1

Table 3: Gender based Model Inversion accuracy.

ID	**Method**	**Init Type**	**Gender (%)**	**Age (%)**	**Accent (%)**	**Time (s)**
1	**GAI** + STD	Zero	71.42 (↑)	18.10 (↓)	6.67 (↓)	23.285
2	**GAI** + STD	Laplacian	65.71 (↑)	26.67 (↑)	11.43 (−)	26.082
3	**GAI** + STD	Uniform	60.95 (↓)	30.48 (−)	11.43 (↓)	27.489

4.5 Qualitative analysis of audio samples

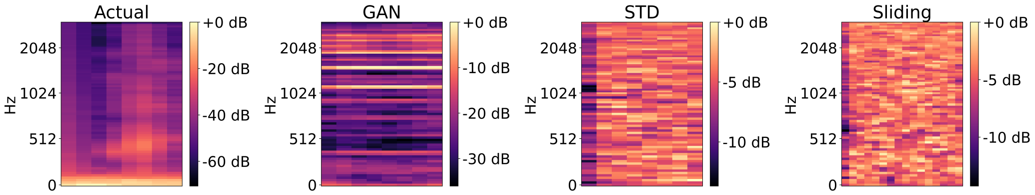


Figure 3: Mel-spectrograms comparison of an actual data point (Actual), a generated mel-spectrogram from the GAN and spectrograms from the model inverted samples

To assess the fidelity and acoustic realism of the reconstructed audio, we performed a qualitative inspection of the spectral characteristics generated by our proposed framework compared to baseline methods. We visualize the Mel-spectrograms of the reconstructed audio samples to analyze their frequency distributions and temporal structures.

Figure 3 presents a side-by-side comparison of the Mel-spectrograms for a target speaker. The figure displays the ground truth audio from the dataset (Actual), our proposed Generative Audio Inversion method (GAN), the standard gradient-descent baselines on raw waveforms STD and SLIDING.

The Actual spectrogram exhibits the characteristic formant structures and harmonic banding typical of human speech, with clear energy concentrations in lower frequencies and temporal silence gaps.

- **GAN (Ours):** As observed in the second panel of Figure 3, the samples generated by our GAI framework demonstrate a higher degree of structural alignment with the ground truth compared to baselines. Crucially, the generative prior enforces a clean spectral envelope, effectively suppressing the high-frequency noise that typically plagues inversion attacks. This indicates that our method successfully produces audio that lies on the manifold of plausible human sounds, though this also contain somewhat artifacts compared to actual.
- **STD (Standard Inversion):** The spectrogram is dominated by broadband noise and scattered energy bursts. It lacks any discernible harmonic structure or temporal coherence, resembling adversarial noise rather than human speech. While this noise might trigger the target class in the model, it is acoustically unintelligible to human listeners.
- **Sliding (SMI):** The fourth panel, representing SLIDING, shows some improvement in temporal variance compared to the standard approach. However, it still suffers from significant fragmentation and high-frequency artifacts (visible as ”static” across the spectrum).

4.6 TIMIT Dataset Experiment

To quantify the importance of training the neural Vocoder (G_{aud}) we use the pretrained HIFIGan without finetuning and only train the mel generator (G_{mel}) for this we utilize the TIMIT dataset Garofolo (1993b) and a speech recognition model trained using the TIMIT dataset. The sincNet model (T) was trained on 16 selected speakers from the origin 462 speakers available in the TIMIT dataset Garofolo (1993b) which will be our $X_{private}$. The speakers were selected such that the we select the speaker from each gender (of 2), each dialect region (of 8) which has the maximum duration of audio.

Table 4: Gender based Model Inversion accuracy.

ID	Method	Init Type	Gender (%)
1	STD	Zero	50.00
2	SLIDING	Zero	50.00
3	GMI (Ours)	Zero	50.00
4	STD	Laplacian	**64.58**
5	SLIDING	Laplacian	60.42
6	GMI (Ours)	Laplacian	52.08
7	STD	Uniform	60.42
8	SLIDING	Uniform	**62.50**
9	GMI (Ours)	Uniform	54.17

As expected and shown by Table 4 when the G_{aud} is not trained in tandem with the G_{mel} the produced audio signals don't seem to capture the biometric features properly. This is probably due the fact of the data shift between the two data generators.

5 Conclusion and Future Work

In this paper, we introduced **Generative Audio Inversion (GAI)**, a novel framework for recovering high-fidelity speaker representations from deep speaker recognition systems. By adapting the concept of Generative Model Inversion to the audio domain, we addressed the limitations of prior optimization-based attacks which typically yield noisy, intelligible waveforms.

Our qualitative analysis demonstrates that leveraging a pre-trained GAN as a distributional prior significantly regularizes the inversion process. Unlike standard gradient descent Wu et al. (2016) or sliding window techniques Pizzi et al. (2022), which struggle to escape high-dimensional noise, GAI successfully reconstructs Mel-spectrograms that exhibit coherent harmonic structures and formant patterns characteristic of natural human speech. This capability not only validates the theoretical vulnerability of models like SincNet but also elevates the practical severity of the threat, as high-fidelity reconstructions are more likely to bypass human verification or serve as seeds for deepfake voice cloning.

5.1 Future Work

While GAI establishes a new baseline for audio model inversion, several avenues remain for future investigation:

Content-Controllable Inversion: Currently, our method reconstructs the speaker's voice using random semantic content derived from the latent space. Future work should explore disentangling speaker identity from linguistic content. Drawing on recent advances in content privacy Williams et al. (2022; 2024), we aim to guide the inversion process to reconstruct specific target phrases, thereby enabling more targeted spoofing attacks.

Defenses and Auditing: As inversion attacks become more sophisticated, robust defenses are required. We plan to investigate whether techniques used in privacy auditing, such as the perturbed loss features proposed for Membership Inference Teixeira et al. (2024), can be adapted to detect or mitigate inversion attempts. Specifically, we aim to explore if training with differential privacy or adversarial regularization can obscure the decision boundaries exploited by our generative search.

Architecture Generalization: This study focused on SincNet. Future experiments will assess the transferability of GAI to more recent architectures, such as ECAPA-TDNN and Conformer-based systems, to determine if the vulnerability is architectural or fundamental to the speaker recognition task.

REFERENCES

Guangke Chen, Yedi Zhang, and Fu Song. Slmia-sr: Speaker-level membership inference attacks against speaker recognition systems. In *Proceedings 2024 Network and Distributed System Security Symposium*, NDSS 2024. Internet Society, 2024. doi: 10.14722/ndss.2024.241323. URL http://dx.dVCTKoi.org/10.14722/ndss.2024.241323.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pp. 1322–1333, New York, NY, USA, 2015a. Association for Computing Machinery. ISBN 9781450338325. doi: 10.1145/2810103.2813677. URL https://doi.org/10.1145/2810103.2813677.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pp. 1322–1333, New York, NY, USA, 2015b. Association for Computing Machinery. ISBN 9781450338325. doi: 10.1145/2810103.2813677. URL https://doi.org/10.1145/2810103.2813677.

J S Garofolo. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. In *Linguistic Data Consortium*. 1993a.

J S Garofolo. Timit acoustic phonetic continuous speech corpus. In *Linguistic Data Consortium*. 1993b.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017. URL https://arxiv.org/abs/1704.00028.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s), September 2022. ISSN 0360-0300. doi: 10.1145/3523273. URL https://doi.org/10.1145/3523273.

Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High Accuracy and High Fidelity Extraction of Neural Networks, a. URL http://arxiv.org/abs/1909.01838.

Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning, b. URL http://arxiv.org/abs/1804.00308.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020. URL https://arxiv.org/abs/2010.05646.

Jacob Leon Kröger, Otto Hans-Martin Lutz, and Philip Raschke. Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference. In Michael Friedewald, Melek Önen, Eva Lievens, Stephan Krenn, and Samuel Fricker (eds.), *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers*, pp. 242–258. Springer International Publishing. ISBN 978-3-030-42504-3. URL https://doi.org/10.1007/978-3-030-42504-3_16.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*, 2019.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 739–753. doi: 10.1109/SP.2019.00065. URL http://arxiv.org/abs/1812.00910.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Fábio Perez and Ian Ribeiro. Ignore Previous Prompt: Attack Techniques For Language Models. URL http://arxiv.org/abs/2211.09527.

Karla Pizzi, Franziska Boenisch, Ugur Sahin, and Konstantin Böttinger. Introducing model inversion attacks on automatic speaker recognition. In *2nd Symposium on Security and Privacy in Speech Communication*, spsc_2022, pp. 11–16. ISCA, September 2022. doi: 10.21437/spsc.2022-3. URL http://dx.doi.org/10.21437/SPSC.2022-3.

Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet, 2019. URL https://arxiv.org/abs/1808.00158.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, 2018. URL https://arxiv.org/abs/1712.05884.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks against Machine Learning Models. URL http://arxiv.org/abs/1610.05820.

Francisco Teixeira, Karla Pizzi, Raphaël Olivier, Alberto Abad, Bhiksha Raj, and Isabel Trancoso. Improving membership inference in asr model auditing with perturbed loss features. *ArXiv*, abs/2405.01207, 2024. URL https://api.semanticscholar.org/CorpusID:269502253.

Kuan-Chieh Wang, given-i=YAN family=FU, given=YAN, Ke Li, Ashish Khisti, Richard Zemel, and Alireza Makhzani. Variational model inversion attacks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9706–9719. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/50a074e6a8da4662ae0a29edde722179-Paper.pdf.

Jennifer Williams, Karla Pizzi, Shuvayanti Das, and Paul-Gauthier Noé. New challenges for content privacy in speech and audio. In *2nd Symposium on Security and Privacy in Speech Communication*, spsc_2022, pp. 1–6. ISCA, September 2022. doi: 10.21437/spsc.2022-1. URL http://dx.doi.org/10.21437/SPSC.2022-1.

Jennifer Williams, Karla Pizzi, Paul-Gauthier Noe, and Sneha Das. Exploratory evaluation of speech content masking, 2024. URL https://arxiv.org/abs/2401.03936.

Xi Wu, Matthew Fredrikson, Somesh Jha, and Jeffrey F. Naughton. A methodology for formalizing model-inversion attacks. In *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, pp. 355–370, 2016. doi: 10.1109/CSF.2016.32.

Yupeng Yin, Xianglong Zhang, Huanle Zhang, Feng Li, Yue Yu, Xiuzhen Cheng, and Pengfei Hu. Ginver: Generative model inversion attacks against collaborative inference. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pp. 2122–2131, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583306. URL https://doi.org/10.1145/3543507.3583306.

Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. URL http://arxiv.org/abs/1911.07135.

A GAI Algorithm

The complete procedure for the proposed Generative Audio Inversion is detailed in Algorithm 1. Unlike standard inversion techniques which perform gradient ascent directly on the raw input space $\mathcal{X}$ Wu et al. (2016), our method optimizes a low-dimensional latent vector z projected through the frozen generative pipeline.

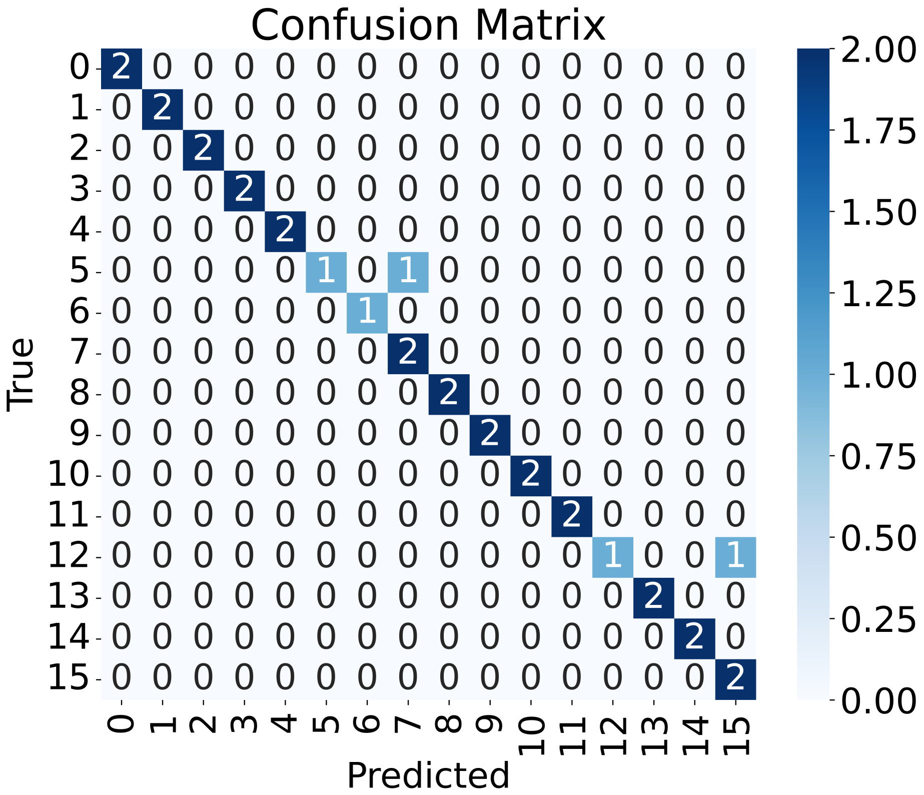
The algorithm begins by initializing z from a standard normal distribution $\mathcal{N}(0, I)$. In each iteration of the optimization loop, the latent vector is transformed into a Mel-spectrogram M and subsequently decoded into a waveform x by the pre-trained vocoder. This candidate audio is then evaluated by the fixed target model T. Gradients are computed with respect to the total loss $\mathcal{L}_{total}$, which balances the adversarial goal of maximizing the target speaker probability ($\mathcal{L}_{id}$) with the constraint of acoustic plausibility ($\mathcal{L}_{prior}$) provided by the discriminators. The latent vector z is updated via gradient descent, effectively traversing the learned manifold to find a representation that is both semantically adversarial and acoustically realistic. n

Algorithm 1 Generative Audio Inversion (GAI) Attack

```
Require:
  Target model $T$ (SincNet)
  Pre-trained Generator $G = \{G_{mel}, G_{aud}\}$
  Target speaker identity $y_t$
  Iterations $\alpha$, Learning rate $\lambda$, Weights $\lambda_{id}, \lambda_{prior}$
Ensure: Reconstructed audio $x^*$
  Freeze parameters of $G_{mel}$ and $G_{aud}$
  Initialize latent vector $z \sim \mathcal{N}(0, I)$
  Initialize optimizer for $z$
  for $i \leftarrow 1$ to $\alpha$ do
      // Forward Pass
      $M \leftarrow G_{mel}(z)$                                  ▷ Generate Mel-spectrogram
      $x \leftarrow G_{aud}(M)$                                  ▷ Vocode to Waveform
      $p \leftarrow T(x)$                                        ▷ Target Model Prediction
      // Compute Losses
      $\mathcal{L}_{id} \leftarrow \mathcal{L}_{CE}(p, y_t)$                     ▷ Cross-Entropy Identity Loss
      $\mathcal{L}_{prior} \leftarrow -(D_{mel}(M) + D_{aud}(x))$              ▷ Adversarial Prior Loss
      $\mathcal{L}_{total} \leftarrow \lambda_{id}\mathcal{L}_{id} + \lambda_{prior}\mathcal{L}_{prior}$
      // Backward Pass (Optimization in Latent Space)
      $z \leftarrow z - \lambda \cdot \nabla_z \mathcal{L}_{total}$                    ▷ Update latent vector
      // Convergence Check (Optional)
      if argmax$(p) == y_t$ and confidence $\geq$ threshold then
          break
      end if
  end for
  $x^* \leftarrow G_{aud}(G_{mel}(z))$
  return $x^*$
```

B TIMIT: Speaker Recognition Model Training.

During the training procedure the model had a validation split of 20% and validation was conducted to check whether model converged. Below displays the validation confusion matrix. The sincnet model was trained with using audio samples of frequecy 16000 Hz and 2s clips (after removing the initial and end of sample pauses). The audio was normalized before training according to specifications in Ravanelli & Bengio (2019)



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2

Figure 4: Validation confusion matrix for the speaker recognition system

C Network Architectures

In this section, we detail the specific architectures used for the Generator (G), the Mel-Spectrogram Discriminator (D_{mel}), and the Audio Discriminator (D_{aud}). All models were implemented in PyTorch. The slope of the leaky ReLU activation was set to 0.2 for all networks.

C.1 Mel-Spectrogram Generator (G_{mel})

The generator takes a latent noise vector z of dimension N (where $N = \texttt{opt.latent\_dim}$) and upsamples it to a Mel-spectrogram of size 80×16. The architecture utilizes a series of upsampling and convolutional blocks.

C.2 Discriminators

We utilize two discriminators: one for the Mel-spectrograms (D_{mel}) and one for the raw audio (D_{aud}). Both utilize a Multi-Layer Perceptron (MLP) architecture.

Table 5: Architecture of the Mel-Spectrogram Generator. The input is a latent vector $z \in \mathbb{R}^N$. The output is a Mel-spectrogram $x \in \mathbb{R}^{80\times16}$.

Layer Type	**Kernel / Stride / Pad**	**Output Shape**	**Activation**
Input (z)	-	N	-
Linear	-	320	-
Reshape	-	$64 \times 5 \times 1$	-
BatchNorm2d	-	$64 \times 5 \times 1$	LeakyReLU (0.2)
Upsample	scale=2	$64 \times 10 \times 2$	-
Conv2d	3×3 / 1 / 1	$128 \times 10 \times 2$	-
BatchNorm2d	-	$128 \times 10 \times 2$	LeakyReLU (0.2)
Upsample	scale=2	$128 \times 20 \times 4$	-
Conv2d	3×3 / 1 / 1	$128 \times 20 \times 4$	-
BatchNorm2d	-	$128 \times 20 \times 4$	LeakyReLU (0.2)
Upsample	scale=2	$128 \times 40 \times 8$	-
Conv2d	3×3 / 1 / 1	$128 \times 40 \times 8$	-
BatchNorm2d	-	$128 \times 40 \times 8$	LeakyReLU (0.2)
Upsample	scale=2	$128 \times 80 \times 16$	-
Conv2d	3×3 / 1 / 1	$64 \times 80 \times 16$	-
BatchNorm2d	-	$64 \times 80 \times 16$	LeakyReLU (0.2)
Conv2d	3×3 / 1 / 1	$1 \times 80 \times 16$	Tanh

Table 6: Architecture of the Mel-Spectrogram Discriminator (D_{mel}). The input is a flattened Mel-spectrogram of dimension $80 \times 16 = 1280$.

Layer Type	**Output Dimension**	**Activation**
Input (Flattened Mel)	1280	-
Linear	512	LeakyReLU (0.2)
Linear	256	LeakyReLU (0.2)
Linear	1	Sigmoid

Table 7: Architecture of the Audio Discriminator (D_{aud}). The input is a raw audio vector of size 4000.

Layer Type	**Output Dimension**	**Activation**
Input (Raw Audio)	4000	-
Linear	1024	LeakyReLU (0.2)
Linear	512	LeakyReLU (0.2)
Linear	256	LeakyReLU (0.2)
Linear	1	Sigmoid