Analysis of Model Inversion Attack and Defense Strategies on Speaker Classification Models: An Architecture-Level Privacy Perspective

Xuan-Yu Lin, Chia-Wei Chuang
National Hualien Senior High School, Hualien, Taiwan

Keywords
Model Inversion Attack; Speaker Recognition; Differential Privacy; Deep Learning Security; Deepfake Voice; SincNet; Waveform Architecture

Article Info
Received: Dec 2025
Revised: March 2026
Accepted: April 2026

Corresponding Author:
Xuan-Yu Lin
National Hualien Senior High School, Hualien, Taiwan
Email: s210347@hlhs.hlc.edu.tw

ABSTRACT

Recent advances in Deepfake voice generation have intensified concerns over speech privacy, as constructing realistic synthetic voices requires access to representative speaker features. Model inversion attacks (MIAs) operationalize this threat by reconstructing such features directly from trained recognition models. However, prior studies rarely examined how model architecture governs inversion vulnerability or under what conditions differential privacy (DP) provides meaningful defense. This work bridges that gap by systematically evaluating the architectural determinants of inversion susceptibility across three speaker classification frameworks—Raw waveform convolutional, SincNet, and MFCC-based models—under white-box gradient-based attacks. Experiments were conducted on three public datasets (50_Speakers, UrbanSound8K, and TIMIT) using both traditional full-gradient and sliding window attack strategies, with performance assessed via attack success rate, attack time, and D-vector reconstruction distance. Key findings demonstrate that: (1) waveform-based architectures provide substantially stronger privacy protection than spectrogram-based models, with the Raw model exhibiting the highest resistance and SincNet offering a favorable trade-off between convergence stability and robustness; (2) a defensive temporal sweet spot exists at approximately 1.6–2.0 seconds of input duration, within which model resistance to inversion is maximized; (3) model depth correlates positively with attack resistance, while the sliding attack decays in effectiveness more rapidly than the full-gradient strategy as model complexity increases; and (4) differential privacy functions as an architecture-dependent defense amplifier—significantly enhancing robustness in Raw and SincNet models while providing negligible benefit to the fundamentally vulnerable MFCC architecture. These findings establish architecture-level design principles for privacy-preserving speech AI systems and demonstrate that effective security requires coherent architectural choices rather than isolated defensive mechanisms.

1. INTRODUCTION

The proliferation of Deepfake voice generation technologies has introduced a new and serious dimension to digital security threats. Highly realistic synthetic voice replication enables fraud, impersonation attacks, and unauthorized access in speaker recognition-based authentication systems [1]. A critical enabler of such attacks is the availability of a target speaker's representative vocal feature embedding—information that may be extracted directly from a deployed speaker recognition model through a technique known as a model inversion attack (MIA). By iteratively optimizing a synthetic input to maximize a model's classification confidence for a target speaker, an attacker can reconstruct vocal characteristics such as timbre, rhythm, and accent without direct access to the original training data [2].

The severity of this vulnerability is amplified by the rapid deployment of speaker recognition systems in commercial and security-critical applications, including voice-activated authentication, forensic speaker identification, and smart device interaction. Despite the well-documented existence of MIAs in image classification domains, their application to speech recognition models—and specifically the role of model architecture in governing inversion susceptibility—remains comparatively understudied. Prior work by Pizzi et al. [2] demonstrated that sliding window MIAs could achieve up to 90% reconstruction accuracy on speaker recognition models, yet did not systematically investigate how architectural choices modulate this vulnerability.

This study addresses that gap by conducting a systematic empirical analysis of how architectural design decisions—specifically the choice of feature extraction front-end, model depth, and input duration—affect model inversion vulnerability in speaker classification systems. In parallel, we evaluate the efficacy of differential privacy (DP) as a post-training defense mechanism, examining how its protective value interacts with the underlying architectural characteristics of the target model.

The central research questions addressed are: (1) How does the choice of feature extraction architecture (Raw waveform, SincNet, MFCC) govern susceptibility to gradient-based inversion attacks? (2) What model depth and input duration configurations optimize the trade-off between classification utility and inversion resistance? (3) Under what architectural conditions does differential privacy provide meaningful defense against model inversion?

Our key findings establish that robustness against inversion is not incidental but is determined by architectural depth and signal representation choices. End-to-end waveform architectures substantially outperform spectrogram-based models in privacy preservation; a defensive temporal sweet spot of approximately 1.6–2.0 seconds of input duration maximizes resistance; and differential privacy functions as an architecture-dependent defense amplifier—effective for inherently robust architectures and largely ineffective for vulnerable ones. These findings contribute an architecture-level framework for the design of privacy-preserving speaker recognition systems.

1.1 Background

1.1.1 Feature Extraction Approaches

Three distinct feature-extraction paradigms were selected as experimental targets to span the spectrum of representation strategies employed in speaker recognition. Mel-Frequency Cepstral Coefficients (MFCC) represent the classical hand-engineered signal processing pipeline, compressing speech into a low-dimensional spectral summary that approximates human auditory perception. This aggressive dimensionality reduction discards substantial fine temporal structure, which—as we demonstrate—critically impacts inversion vulnerability. SincNet [1] is an end-to-end architecture that retains signal processing priors through parameterized sinc filters in its first layer, learning optimal low- and high-frequency cutoff boundaries while preserving substantially more waveform detail than MFCC. The Raw waveform model applies standard convolutional filters directly to the raw audio signal without any pre-filtering, providing a maximum-capacity, high-fidelity representation that retains the richest low-level acoustic structure.

1.1.2 Model Inversion Attack Taxonomy

Model inversion attacks can be distinguished from related adversarial threat classes by their objective: rather than forcing misclassification (adversarial attacks) or determining training set membership (membership inference attacks), MIAs reconstruct training-like inputs directly from model internals. This produces concrete, human-interpretable reconstructions and therefore constitutes a severe speech privacy threat. All experiments in this study employ a white-box attack setting, in which full parameter and gradient access is available. While this represents an upper bound on attacker capability, it provides controlled, reproducible comparisons across architectures and is the most informative setting for vulnerability analysis.

1.1.3 Differential Privacy as Defense

Differential Privacy (DP) is a formal mathematical framework that provides privacy guarantees by ensuring that small changes in the training dataset do not significantly alter model outputs. In deep learning, it is conventionally implemented by injecting calibrated noise into training gradients—a computationally expensive approach that often destabilizes convergence for high-dimensional audio models. This study adopts an alternative post-training DP strategy inspired by the One Parameter Defense framework [3], which appends a lightweight DP layer to the model's softmax output. This layer applies an exponential mechanism that perturbs output probabilities while preserving their relative ranking, injecting output uncertainty that misleads gradient-based attackers while maintaining classification accuracy. Figure 1 illustrates the DP layer architecture.

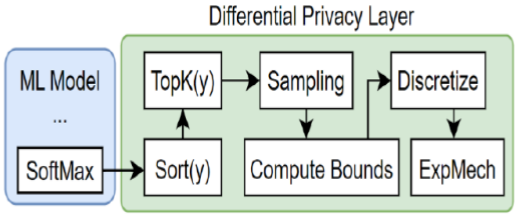

Figure 1. Architecture of the post-training Differential Privacy layer, implementing the exponential mechanism on top-K softmax outputs via sampling, bounds computation, and discretization.

1.2 Related Work

The foundational work of Ravanelli and Bengio [1] introduced SincNet, demonstrating that learnable sinc-filter front-ends improve convergence and interpretability in speaker recognition from raw waveforms. This architecture forms one of the primary experimental targets in this study. Pizzi et al. [2] introduced model inversion attacks on automatic speaker recognition, proposing a Sliding Model Inversion Attack that outperformed the traditional full-gradient method—achieving 90% versus 54% reconstruction accuracy on the tested architecture. Our work critically extends this finding by demonstrating that the sliding attack's superiority is architecturally conditional, a key qualification not established by the original study. The One Parameter Defense framework [3], which proposed a simplified output-level DP mechanism for image classification, informed the DP defense implementation adapted for speech models in this study.

2. RESEARCH METHOD

2.1 Methodology Overview

The experimental methodology consists of three sequential stages: (1) dataset processing and standardization; (2) target model training across varying architectures, complexities, and DP configurations; and (3) model inversion attack execution and evaluation. Figure 2 provides a schematic overview of the complete experimental pipeline. Five controlled variables are systematically varied to isolate their individual contributions to inversion vulnerability (Table 1).

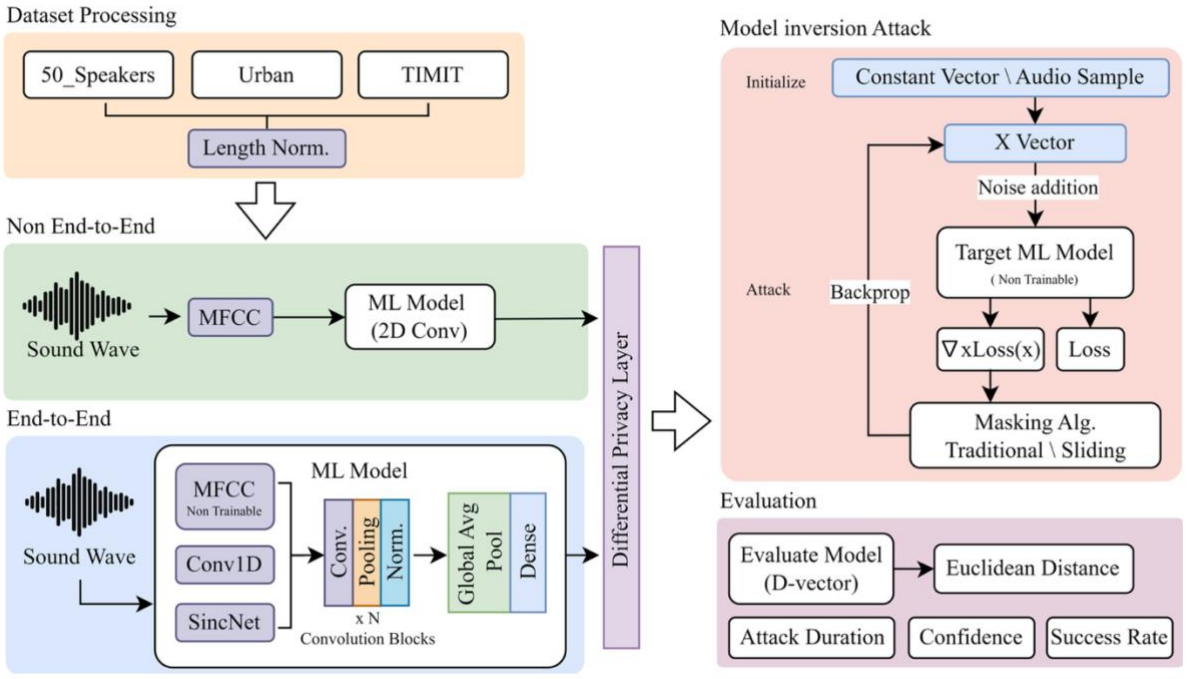

Figure 2. Complete experimental pipeline: dataset processing (left), end-to-end and non-end-to-end model architectures with optional Differential Privacy layer (center), and model inversion attack with evaluation metrics (right).

Table 1. Controlled Variables Across Experimental Stages

Experimental Stage	Controlled Variable	Description
Dataset Processing	Sampling length	Audio segment duration ranging from 3,200 to 51,200 samples (≈0.2–3.2 s at 16 kHz), tested in nine discrete steps
Target Model Training	Architecture & input feature	Three configurations: Raw (end-to-end waveform), SincNet (learnable filter front-end), and Non-E2E MFCC (2D conv on pre-computed features)
	Model complexity	Convolutional depth scaled from Structure 0 (4 blocks) to Structure 4 (8 blocks); channel width also varied
	Differential Privacy defense	Binary: DP post-processing layer applied or not; implemented via exponential mechanism on softmax output
Model Inversion Attack	Seed initialization	Real audio sample (dataset-sourced) vs. synthetic initializations (zeros, ones, constants, random noise)

2.2 Dataset Processing

All audio recordings were resampled to 16 kHz. Three public datasets were utilized to ensure experimental generalizability across both speech and non-speech acoustic domains: 50_Speakers [4], UrbanSound8K [5], and TIMIT [6] (Table 3). The inclusion of UrbanSound8K—a non-speech environmental sound dataset—alongside the speech-centric 50_Speakers and TIMIT corpora enables assessment of model behavior under both vocal and non-vocal acoustic conditions.

Table 2. Dataset Composition and Selection Policy

Dataset	Classes	Clips per Class (Selection Rule)
50 Speakers [4]	50	1,000 randomly selected per class (all included if < 1,000)
UrbanSound8K [5]	10	1,000 randomly selected per class (all included if < 1,000)
TIMIT [6]	463	200 randomly selected per class

Input duration was controlled as the primary variable in the temporal analysis experiment. Audio segments were extracted using a sliding window approach, with nine discrete lengths tested: 3,200, 6,400, 12,800, 19,200, 25,600, 32,000, 38,400, 44,800, and 51,200 samples (corresponding to approximately 0.2–3.2 s at 16 kHz). A segment was retained only if more than 50% of its duration contained speech-active content, defined as non-silent and semantically meaningful acoustic information. All retained segments were trimmed or zero-padded to the target length. The primary working length for architecture comparison experiments was 3,200 samples (≈ 0.2 s); model complexity experiments used 64,000 samples as the standard configuration.

2.3 Target Model Training

Three model architecture families were evaluated. Raw and SincNet are end-to-end (E2E) architectures that operate directly on raw waveform inputs; a third E2E variant incorporates a non-trainable MFCC front-end layer feeding 1D convolutional blocks. The Non-E2E baseline uses pre-computed MFCC features as input to a 2D convolutional network. All models share a common backbone consisting of four convolutional blocks in the baseline configuration (Structure 0), each block following a Convolution → Pooling → Normalization sequence. Model depth was extended from Structure 0 (4 blocks) to Structure 4 (8 blocks) in the complexity analysis. Feature sizes across architectures are compared in Table 3.

Table 3. Feature Dimensionality of Evaluated Architectures

Architecture	Feature Length (Samples)	Feature Dimensions	Total Feature Size
Raw (waveform)	3,072	80	245,760
MFCC	350	1 (1D)	350
SincNet	200	80	16,000

The Differential Privacy defense was implemented as a binary experimental condition—applied or not applied—at the model output stage. The DP layer modifies the softmax output via the exponential mechanism, preserving the relative ranking of class probabilities while injecting calibrated noise that increases attack uncertainty without degrading classification accuracy.

2.4 Model Inversion Attack and Evaluation

For each target class, an input signal was optimized through iterative gradient descent to minimize the cross-entropy loss between the model's softmax output and a one-hot target vector. The attack terminates when this loss falls below a threshold of 0.005, or after a maximum of 10,000 iterations. Two attack strategies were employed:

Traditional Full-Gradient Attack: The gradient is computed across the entire input sample at each iteration, with the full sample updated via backpropagation. This strategy targets globally integrated feature representations.

Sliding Window Attack: Gradients are computed and applied to short, overlapping temporal windows sequentially. Two optimized parameter sets were used: (i) window=430, step=85 (highest attack confidence); and (ii) window=410, step=100 (shortest attack time). Results from both configurations were averaged for robust reporting.

Three primary evaluation metrics were applied. Attack Success Rate: an attack was deemed successful if the reconstructed audio achieved model confidence exceeding 80% for the target class; success rate is reported as the percentage of target classes successfully attacked. Attack Time: total elapsed computation time for successful attack completion. D-Vector Reconstruction Distance: a pre-trained generalized speaker embedding model was used to extract D-vectors from both the reconstructed and original audio samples; the Euclidean distance between the synthesized D-vector and the centroid of original class D-vectors quantifies reconstruction fidelity (lower distance = higher information leakage).

2.5 Implementation Environment

All experiments were conducted on an Apple MacBook Pro with Apple M2 Max chip (12-core CPU, 38-core GPU, 32 GB unified memory). Model training and inversion procedures were implemented in Python 3.9 using TensorFlow 2.13, executed locally without external GPU clusters.

3. RESULTS AND DISCUSSION

3.1 Effect of Audio Clip Length

The relationship between input duration and attack performance reveals a non-monotonic pattern with significant implications for defensive design (Figure 3). For the sliding window attack, shorter clips yielded markedly lower attack time—a straightforward consequence of reduced input dimensionality. The traditional full-gradient attack showed a steady increase in attack cost with clip length, consistent with the growing complexity of optimizing over a longer input vector.

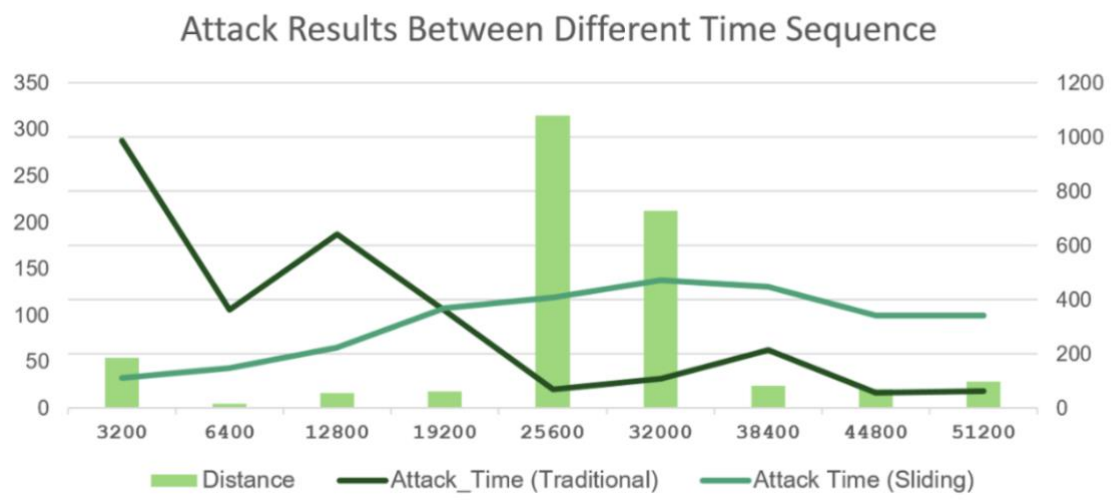

Figure 3. Attack performance as a function of audio clip length (samples): D-vector distance (bars, left axis) and attack time for traditional and sliding methods (lines, right axis). A distinct Resistance Zone is observed at approximately 25,600–32,000 samples (1.6–2.0 s).

Critically, a distinct Resistance Zone was identified at approximately 25,600–32,000 samples (1.6–2.0 seconds). Within this range, the sliding attack incurred its highest computational cost and produced reconstructions with the greatest D-vector distance from the original training samples—indicating minimal information leakage. This phenomenon suggests that at this temporal scale, the model's internal feature encoding achieves a particularly stable and globally integrated representation, making localized perturbations by the sliding attack insufficient to capture the classifier's decision boundary. This finding has practical implications for defense: configuring input segment duration within this range constitutes a low-cost, architecture-agnostic defensive measure.

3.2 Impact of Architecture: End-to-End vs. Non-End-to-End

The choice between end-to-end (waveform-based) and non-end-to-end (spectrogram-based) architectures produced one of the most consequential differences observed in this study (Figure 4). Attacks on spectrogram-based models (MFCC Non-E2E) yielded substantially smaller D-vector distances than attacks on waveform-based models, indicating higher reconstruction fidelity and therefore greater information leakage. The richer time-frequency structure captured in MFCC representations—while beneficial for classification—paradoxically creates a more exploitable gradient surface for attackers.

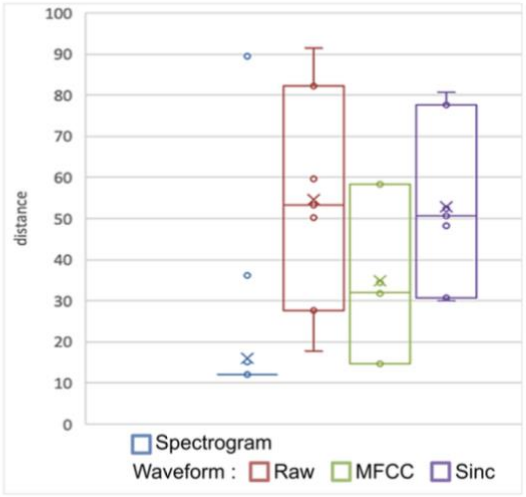

Figure 4. D-vector distance between reconstructed and original audio for spectrogram-based (Spectrogram) and waveform-based (Raw, MFCC E2E, SincNet) architectures. Smaller distance indicates more accurate reconstruction and greater privacy leakage.

Waveform-based architectures require the model to internally derive complex, high-dimensional feature representations from raw signal structure—a process that is inherently difficult to reverse through gradient optimization. This provides a meaningful baseline of inversion resistance that cannot be easily replicated by adding post-hoc defenses to a spectrogram-based model.

3.3 Influence of Feature Extraction Architecture

Among end-to-end architectures, the feature extraction strategy produces significant and interpretable differences in both attack time and reconstruction quality (Figure 5; Table 2). The Raw model, with its feature space of 245,760 dimensions, imposed substantially longer attack times—reflecting the computational burden of optimizing over a high-dimensional input space. Correspondingly, reconstructed samples from Raw models exhibited the largest D-vector distances from original training data, confirming that high-dimensional feature preservation constitutes a genuine defensive advantage.

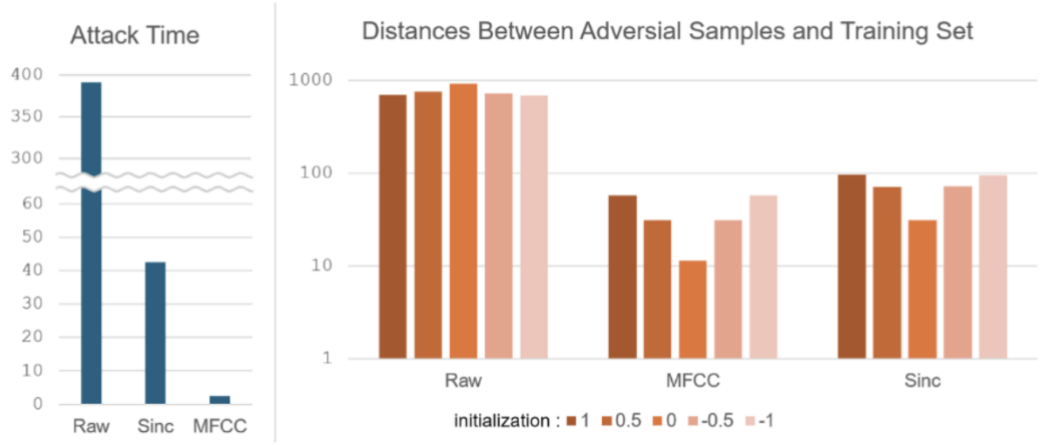

Figure 5. Comparison of attack time (left) and D-vector distance between adversarial samples and training set (right, log scale) across Raw, MFCC, and SincNet architectures under different initialization strategies.

MFCC and SincNet architectures, by applying early-stage dimensionality reduction (to 350 and 16,000 features respectively), significantly reduced attack time by constraining the optimization problem to a lower-dimensional space. However, this efficiency came at the cost of producing reconstructions more closely resembling original training samples—evidenced by smaller D-vector distances. The trade-off between feature compactness and privacy resistance is thus quantified: MFCC's extreme compression (350 features) produces the fastest attacks and greatest information leakage, while SincNet's intermediate dimensionality (16,000 features) and learnable filter design offer a practical balance between training stability, classification performance, and inversion resistance.

3.4 Effect of Initialization Strategy

The initialization strategy for the attack seed had a counterintuitive but consistent effect on attack performance (Figure 6). Initializing the attack from a real audio sample (e.g., drawn from UrbanSound or TIMIT datasets) consistently produced higher attack costs than synthetic initializations (zeros, mean values, random noise). This observation indicates that genuine audio structures—with their complex temporal correlations and energy distributions—create a more difficult starting point for gradient optimization, as the inherent structure of the real signal resists modification into a class-representative adversarial example without introducing highly visible distortions.

Table 4. Feature Size of Different Feature Extraction Architectures

	Length	Dimensions	Feature Size
Raw	3072	80	245760
MFCC	350	1	350
SincNet	200	80	16000

Synthetic initializations, particularly mean-value and random noise seeds, provide a neutral, unstructured starting point that the gradient descent process can more readily sculpt into a target-class representation. This finding suggests that real-audio seed initialization inadvertently reduces attack efficiency—a potentially exploitable insight for defensive seed strategies.

3.5 Impact of Model Complexity

Model depth shows a direct and consistent positive relationship with attack difficulty (Figure 7). As architectural depth increases from Structure 0 (4 convolutional blocks) to Structure 4 (8 blocks), attack time rises and both success rate and prediction confidence decline—confirming that deeper models are inherently more resistant to inversion. This relationship supports the theoretical expectation that deeper architectures transform local acoustic features into progressively more abstract, globally integrated representations that are harder to invert.

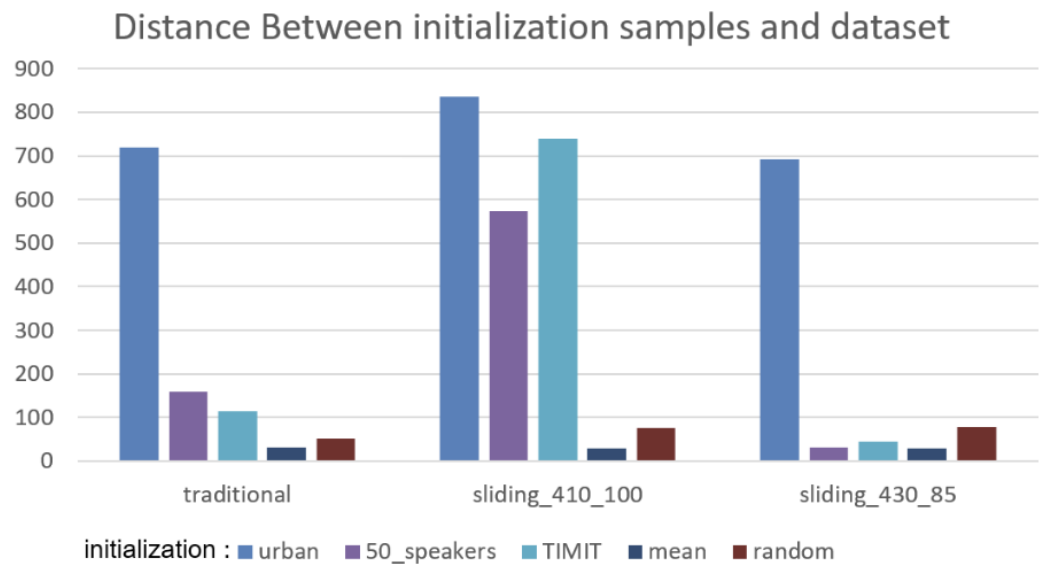

Figure 6. Performance of traditional and sliding attacks across model complexity levels (Structure 0 to Structure 4): attack time (s, bars), D-vector distance (scaled, hatched bars), and success rate (lines).

A critical asymmetry was also observed: the sliding window attack's performance degraded significantly faster than the traditional full-gradient attack as model depth increased. This finding provides an important qualification to the conclusions of Pizzi et al. [2], who reported sliding attack superiority without controlling for architectural depth. In the framework proposed in Section 3.6, this observation is explained by the locality of the sliding attack, which exploits shallow, locally retained features that are progressively abstracted away by deeper layers. The full-gradient attack, targeting the model's global output representation, retains utility even as local features become inaccessible.

3.6 Evaluation of Differential Privacy as Defense

The effectiveness of the DP defense layer showed strong architectural dependence, confirming the central hypothesis of this study (Figure 8). For Raw and SincNet architectures, DP integration produced substantial increases in attack time and marked reductions in attack success rate. For the MFCC architecture, the improvement was largely negligible.

Figure 7. Effectiveness of Differential Privacy defense across architectures (Raw, SincNet, MFCC) and attack strategies: attack time with/without DP (bars, log scale, left axis) and attack success rate with/without DP (lines, right axis).

This architecture-dependent pattern reveals that DP cannot compensate for fundamental representational vulnerabilities. For Raw and SincNet models, DP's gradient-obfuscating effect operates within a high-dimensional, complex feature space, substantially increasing the attacker's optimization burden. For MFCC models, the front-end bottleneck reduces the feature space to 350 dimensions before any gradients are computed, making it straightforward for an attacker to operate within this simplified space regardless of back-end output perturbation. DP in this context functions analogously to fortifying the interior of a structure that lacks a perimeter—attackers bypass the defense by operating within the unprotected front-end domain.

3.7 Theoretical Framework: Hierarchical Information Processing

To provide a unified theoretical account of the experimental observations, we propose the Hierarchical Information Processing Framework. In this model, a deep neural network functions as a multi-level information filter: raw input audio contains the richest, most locally structured information, which is progressively compressed and abstracted as it propagates through successive network layers. Each layer discards features deemed irrelevant to the classification task while consolidating retained features into more globally integrated representations. The transition from local to global representation is the central determinant of inversion vulnerability.

This framework explains the observed attack strategy asymmetry: the sliding window attack exploits local features that are characteristic of shallow processing, becoming less effective as global abstraction increases with model depth; the full-gradient attack targets the model's final global representation, retaining effectiveness in deeper architectures. It also explains the MFCC vulnerability: the fixed, non-adaptive front-end bottleneck pre-compresses information before any learning occurs, creating a persistent, easily exploitable entry point that architectural depth cannot overcome. Finally, it accounts for DP's architecture-dependence: gradient obfuscation is most effective when the attacker must navigate a complex, high-dimensional feature landscape—a condition satisfied by Raw and SincNet models but not by MFCC.

A summary of all experimental findings within this framework is provided in Table 5.

Table 5. Summary of Key Experimental Findings and Theoretical Interpretations

Experimental Factor	Key Metric Outcome	Interpretation
Audio clip length (samples)	Resistance Zone: 25,600–32,000 samples (1.6–2.0 s)	Sliding attack cost peaks; D-vector distance maximized; model encoding exhibits localized temporal robustness
Architecture: E2E vs Non-E2E	D-vector distance: Spec < Waveform (smaller = more leakage)	Spectrogram-based models leak more information; waveform models are inherently more resistant

Experimental Factor	Key Metric Outcome	Interpretation
Feature extraction (Raw, MFCC, SincNet)	Attack time: Raw >> Sinc > MFCC; Distance: Raw >> Sinc > MFCC	High-dimensional feature spaces (Raw) resist inversion; MFCC's low-dimensional bottleneck enables rapid reconstruction
Initialization strategy	Real audio seed → higher attack cost than synthetic	Genuine audio structures impede gradient optimization; synthetic initializations converge more readily
Model complexity (struc 0–4)	Sliding success rate decays faster than traditional as depth increases	Deep models favor full-gradient attacks; locality of sliding attack loses advantage with abstraction depth
Differential Privacy defense	Raw/SincNet: large success rate reduction; MFCC: negligible effect	DP amplifies pre-existing robustness; ineffective where front-end bottleneck is the primary vulnerability

4. CONCLUSIONS

This study systematically examined how architectural design decisions govern the susceptibility of speaker classification models to model inversion attacks, and identified the conditions under which differential privacy provides meaningful defense. The following principal conclusions are drawn:

(1) Architecture Determines Baseline Privacy: End-to-end waveform architectures (Raw, SincNet) provide substantially stronger inversion resistance than spectrogram-based models (MFCC Non-E2E). The Raw model offers the highest resistance through its high-dimensional, learnable feature space; SincNet provides a practical trade-off between convergence stability and defensive strength. MFCC architectures impose a fixed, low-dimensional front-end bottleneck that renders them fundamentally vulnerable regardless of subsequent defensive measures.

(2) Temporal Configuration Matters: A defensive sweet spot at approximately 1.6–2.0 seconds of input duration maximizes inversion resistance, as evidenced by peak attack cost and maximum D-vector distance. Input duration should be treated as a controllable privacy parameter in speaker recognition system design.

(3) Depth Enhances Robustness Asymmetrically: Increasing model depth improves resistance for both attack strategies, but the sliding attack's effectiveness degrades faster than the full-gradient attack with increasing depth. This constrains the conditions under which sliding attacks offer the advantage reported by Pizzi et al. [2] to shallow, simple architectures.

(4) Differential Privacy is an Architecture-Dependent Amplifier: DP should be viewed as a mechanism that amplifies pre-existing architectural robustness rather than as a universal countermeasure. It provides substantial benefit to Raw and SincNet architectures and negligible benefit to MFCC, confirming that effective defense requires coherent architectural choices as the foundation.

For practitioners designing privacy-resilient speaker recognition systems, the implications are clear: adopt deep, waveform-based architectures (Raw or SincNet); configure input segment duration within the 1.6–2.0 second range; and deploy differential privacy as a defense amplifier atop an already robust foundation. Future research directions include: exploration of GAN-based generative attack frameworks capable of producing perceptually realistic speech reconstructions; formal privacy-utility trade-off analysis across a broader range of architectures; and evaluation of combined defenses (architectural + DP + input perturbation) in realistic threat models.

ACKNOWLEDGEMENTS

The authors express their sincere gratitude to Mr. Yi-Hsiung Chao and Professor Greg Lee for their invaluable guidance, insightful advice, and continuous encouragement throughout the development of this research. Their expertise in computer science and machine learning greatly shaped the design and analytical rigor

of this study. The authors also extend sincere thanks to the Department of Science and Mathematics Education, National Hualien Senior High School, for providing research facilities and technical resources. Heartfelt appreciation is given to classmates and family members for their constant support and understanding throughout the experimentation and writing process.

REFERENCES

[1] Ravanelli, M., & Bengio, Y. (2019). Speaker recognition from raw waveform with SincNet. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 1021–1028). IEEE. https://doi.org/10.1109/SLT.2018.8639585

[2] Pizzi, K., Boenisch, F., Sahin, U., & Böttinger, K. (2022). Introducing model inversion attacks on automatic speaker recognition. In Proceedings of the 2nd Symposium on Security and Privacy in Speech Communication (pp. 11–16). https://doi.org/10.21437/SPSC.2022-3

[3] Ye, D., Shen, S., Zhu, T., Liu, B., & Zhou, W. (2022). One parameter defense: Defending against data inference attacks via differential privacy. IEEE Transactions on Information Forensics and Security, 17, 1466–1480. https://doi.org/10.1109/TIFS.2022.3163591

[4] Jain, V. (2019). Speaker recognition audio dataset. Kaggle. https://www.kaggle.com/datasets/vjcalling/speaker-recognition-audio-dataset

[5] Salamon, J., Jacoby, C., & Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In Proceedings of the 22nd ACM International Conference on Multimedia (pp. 1041–1044). https://doi.org/10.1145/2647868.2655045

[6] Fekadu, M. (2019). DARPA TIMIT acoustic-phonetic continuous speech corpus. Kaggle. https://www.kaggle.com/datasets/mfekadu/darpa-timit-acousticphonetic-continuous-speech

[7] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference* (pp. 265–284). Springer. https://doi.org/10.1007/11681878_14

[8] Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (pp. 1322–1333). https://doi.org/10.1145/2810103.2813677

[9] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 308–318). https://doi.org/10.1145/2976749.2978318

[10] Variani, E., Lei, X., McDermott, E., Moreno, I. L., & Gonzalvo, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4052–4056). https://doi.org/10.1109/ICASSP.2014.6854363

[11] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5329–5333). https://doi.org/10.1109/ICASSP.2018.8461375

[12] Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. In *Proceedings of Interspeech 2017* (pp. 2616–2620). https://doi.org/10.21437/Interspeech.2017-950

[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (Vol. 27). Curran Associates.

[14] Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., & Song, D. (2020). The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 253–261). https://doi.org/10.1109/CVPR42600.2020.00033

BIOGRAPHIES OF AUTHORS

Xuan-Yu Lin is a student researcher at National Hualien Senior High School, Hualien, Taiwan. His research interests encompass model inversion attacks, privacy-preserving architectures for speaker recognition, natural language processing, deep learning, and information security. He aims to pursue advanced studies in the interdisciplinary areas of artificial intelligence and electrical engineering.

Chia-Wei Chuang is a student at National Hualien Senior High School, Hualien, Taiwan. His research focuses on model inversion attacks and defense architectures for speaker recognition systems. His broader interests include language processing, machine learning, and related engineering disciplines. He has also contributed to the development of multimodal AI communication systems designed to assist patients with aphasia. He plans to continue his academic journey in artificial intelligence and electrical engineering.