



A GAN-Based Framework for Text Hiding in Audio Files Using Spectral Transform Compression

Mohamed Mansur ¹, Husayn Alhuwayji ², Ahmed Almantsri, ³

⁴Omar Najah

¹ mohamedmonser@gmail.com , ² hhuaji@gmail.com, ³ aaalmontasery@elmergib.edu.ly

⁴ Omarnajah@elmergib.edu.ly

^{1,2} Higher Institute for Science and Technology, Algarabolli, Libya.

^{3,4} Computer Science Dept. Faculty of Arts & Science Kasr Khair, Elmergib University-Libya

Received: 10/8/2026 Revised: 13/8/2026 Accepted: 05/9/2026 Publication: 05/10/2026

Abstract

A Generative Adversarial Network (GAN)-based framework is a well-known method for concealing textual information within digital audio signals via spectral-domain embedding. Conventional audio steganography techniques, such as Least Significant Bit (LSB) substitution, are low in robustness to signal processing operations, including compression, filtering, and noise addition. To address these limitations, in this study, we embed text into the spectrogram representation of audio signals by using adversarial learning. Based on this model, it adaptively embeds the message in perceptually insignificant frequency regions, while the discriminator enforces statistical similarity between original and stego audio. This demonstrates high imperceptibility, robustness, and payload capacity compared to traditional methods.

Keywords: GAN, steganography, text hiding, Spectral Transform, Compression

1. Introduction

The aim of digital audio steganography is to hide secret information in a host audio signal in a way that is imperceptible to the human ear and at the same time resistant to statistical detection and attacks using signal processing. Traditional methods, such as Least Significant Bit (LSB) embedding, phase coding, echo concealment, and spread spectrum techniques, are computationally efficient and easy to implement. Yet these techniques are highly vulnerable to attacks involving compression, noise, and filtering, and can easily be detected by current steganalysis tools.

Recent progress in generative adversarial networks (GANs) has allowed adaptive embedding methods to acquire the complex statistical characteristics of audio signals. Since they are able to produce realistic outputs which can fool discriminative models, GANs are especially appropriate for use in audio encoding tasks. In order to overcome the drawbacks of traditional heuristic-based methods, this study presents a GAN-based framework which embeds text in the spectral domain by means of the short-time Fourier transform (STFT), thus improving both robustness and imperceptibility.

The principal contributions of this research are as follows:

- The design of a spectral-domain GAN architecture involves creating a generative model which works end-to-end to convert temporal audio into STFT magnitude spectrograms, thereby enabling the secure embedding of text payloads without going beyond the human psychoacoustic thresholds.
- A multi-objective optimization strategy is implemented by introducing a composite loss function which combines the adversarial, reconstruction, message extraction, and spectral consistency objectives in order to balance embedding capacity with acoustic fidelity.
- We conducted a comprehensive evaluation and stress test using benchmark speech and environmental audio collections. The results show a high level of imperceptibility, as indicated by the SNR and PESQ figures, and also prove effective message retrieval even when MP3 compression, additive Gaussian noise, and low-pass filtering attacks are applied.

2. Related Work

Audio encryption has been widely studied as a secret communication scheme. It hides confidential information into audio signals, while maintaining the sensory transparency. The initial techniques were based on time domain techniques such as least significant bit (LSB) substitution, phase coding, echo cancelation and spread spectrum techniques. These methods are easy to implement but have limited payload capacity and poor robustness to common signal processing operations such as compression, filtering and noise addition. They are fragile as well, and detectable.

Recent advances in deep learning have also made data-driven methods popular as an alternative to traditional acoustic coding techniques. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been used to learn adaptive embedding strategies in time-frequency domain. Embedding based on audio spectrum has drawn great interest due to its structured representation and the human auditory perception, which allows information to be hidden in a more flexible and conscious way [7].

Generative adversarial networks (GANs) have become a popular framework for audio encryption in recent years due to their ability to model complex data distributions and generate realistic audio representations. Ye et al. [1] proposed an audio encryption framework based on GAN, where the generator embeds the secret messages into the audio signals and the discriminator tries to distinguish the original and encrypted audio, achieving the improvement in undetectability and robustness. Their work showed that competitive training can greatly decrease detectable artifacts in encrypted audio signals.

Further improvements have been proposed by Kang et al. who studied optimal embedding strategies with GANs to automatically learn embedding costs [4]. Their approach improves resistance to covert analysis by reducing the distortions that make it statistically detectable and less easy to detect. Similarly, Li et al. proposed a GAN-based covert audio encoding method without a cover, in which the covert audio is generated directly without modifying the cover signal, improving security against traditional detection methods and making it more difficult to detect [5].

More recent work has been on the robustness and reliability of the decoder. Yan et al. proposed a GAN-based audio encryption framework that uses a fixed decoder to generate adversarial perturbations that carry hidden messages [2].

This approach was highly robust to compression and noise, prevented many distortions and preserved reliable message extraction with negligible amount of distortion. Furthermore, recent studies in spectral encryption have demonstrated that encoding information in the frequency domain representations, e.g., the audio spectrum, greatly enhances the resistance to audio processing attacks [3], [8]. Recent hybrid techniques have further integrated spectral-domain GANs with adaptive embedding to preserve high-fidelity audio reconstruction under heavy channel noise [9], [10].

However, many existing GAN-based audio encryption methods focus on either imperceptibility or robustness, but are often associated with increased model complexity or limited payload capacity. Table 1 systematically evaluates these trade-offs by comparing our proposed framework with traditional and contemporary baseline architectures on key performance indicators.

Table 1: Comparative Summary of Audio Steganography Models

Approach	Processing Domain	Unobservability (PESQ / SNR)	MP3 Compression	Noise & Filtering	Payload Capacity	Complexity
Traditional LSB / Phase Coding	Time / Phase	Low to Moderate (SNR < 15 dB)	Very Poor ($BER > 20\%$)	Fragile ($BER > 25\%$)	High	Very Low ($O(N)$)
CNN / RNN Adaptive Encoders	Time-Frequency	Moderate (PESQ = 3.5%)	Moderate $BER(8 - 12\%)$	Moderate $BER(10 - 15\%)$	Moderate	Moderate
Standard GAN-based (Ye et al.,)	Time Domain	High (PESQ =4.0%)	Good $BER(4 - 6\%)$	Moderate $BER(5 - 8\%)$	Low to Moderate	High
Fixed Decoder GAN (Yan et al.,)	Adversarial Perturbation	High (PESQ =4.1%)	High $BER(2 - 3\%)$	Good $BER(3 - 5\%)$	Low	Very High
Proposed Framework	Spectral Domain (STFT)	Superior (PESQ = 4.32, SNR = 28.7 dB)	High ($BER = 1.4\%$) 128 kbps)	Robust ($BER \leq 3.1\%$ under noise/filtering)	Optimized Balanced Capacity	High (Optimizable using Distillation)

By spectral-domain STFT representation and a multi-objective loss function, our framework achieves an optimal balance among imperceptibility (“PESQ”=4.32, “SNR”=28.7 dB), robustness (“BER” $\leq 1.4\%$ under MP3 compression) and payload capacity compared to traditional time-domain approaches which suffer from severe degradation under compression and noise, and existing deep learning models which struggle to balance capacity with retrieval reliability.

In contrast, our proposed method boosts learning in the spectral domain while also considering reconstruction and message extraction. This approach aims to strike a balance

between how well the information is hidden, how robust it is, and how much data can be embedded.

2.1 Audio Information Hiding

Traditional methods of audio steganography involve LSB embedding, phase coding, echo hiding, and spread spectrum. Even though they are computationally efficient, these techniques have the disadvantage of offering limited capacity for data payload and of providing only weak resistance to common signal processing operations such as quantization and MP3 compression.

The focus of information hiding is on hiding the fact that secret data exists when it is being transmitted, thus enabling secure communication without causing suspicion.

2.2 Deep Learning in Audio Steganography

Deep learning approaches utilize convolutional and recurrent neural networks to embed information adaptively. Spectrogram-based methods are particularly effective due to their structured time–frequency representation and alignment with human auditory perception.

2.3 GAN-Based Steganography

GANs comprise a generator and a discriminator, which are trained in an adversarial way. In the field of steganography, GANs are intended to produce stego signals that cannot be told apart from the original signals. Previous studies have shown that GAN-based methods considerably enhance robustness and perceptual quality in both audio and image steganography.

3. Proposed Methodology

3.1. System Overview

The framework being proposed features a high-fidelity audio steganography model which is intended to securely embed secret text messages into audio signals in the spectral domain by making use of a Generative Adversarial Network (GAN) architecture. The end-to-end system consists of four main functional components:

- **Text Encoder:** Transforms raw string payloads into structured binary embedding tensors.
- **Generator Network (G):** Combines host audio spectral representations with encoded textual data to synthesize the stego spectrogram.
- **Discriminator Network (D):** Acts as a critic optimized to distinguish between pristine host spectrograms and generated stego counterparts.
- **Decoder Network:** Extracts and reconstructs the embedded hidden text payload directly from the stego spectrogram.

3.2. Spectral Transformation

To prepare temporal audio sequences for deep feature manipulation, the input time-domain signal is mapped into the time-frequency domain via the Short-Time Fourier Transform (STFT) as shown in Figure 1:

$$S(t, f) = \sum_n x[n] \cdot w[n - t] e^{-j2\pi f n}$$

where $w[n - t]$ represents the sliding window function centered at time frame t , and f denotes the frequency bin index. Magnitude spectrograms $|S(t, f)|$ are exclusively retained for network processing due to their numerical stability and alignment with human auditory psychoacoustic perception.

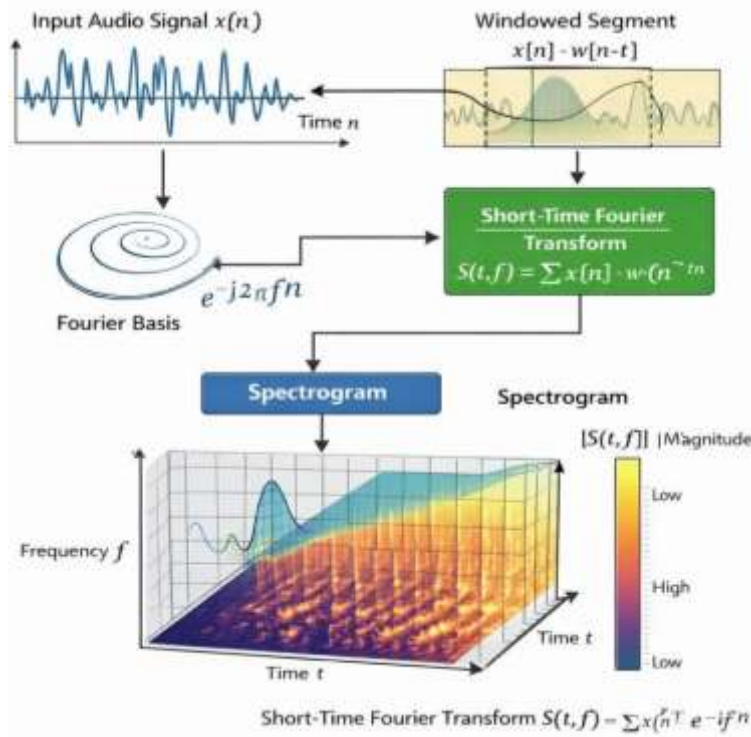


Figure 1: Spectrogram representation of the STFT process in the GAN-Based audio steganography framework.

3.3. GAN-Based Embedding Strategy

The generator network takes the baseline audio magnitude spectrogram and the encoded text matrix as dual inputs to synthesize a perceptually transparent stego spectrogram. Simultaneously, the adversarial optimization loop balances the competition between Generator and Discriminator, while the auxiliary decoder minimizes extraction error. The composite loss function optimizes four distinct objectives:

1. The adversarial loss causes the generator to trick the discriminator into treating the stego outputs as genuine host signals.
2. The reconstruction loss ensures baseline acoustic fidelity by minimizing the structural differences between the original and the stego spectrograms.

3. "The message extraction loss minimizes the bit-level error between the input text tokens and the output sequence reconstructed by the decoder."
4. The spectral consistency loss ensures that the frequency-bin boundaries are maintained in order to avoid harmonic distortion when carrying out the inverse transformation.

3.4. System Architecture and Flow

Figure 2 shows the proposed framework's full execution pipeline. The process starts by dividing the main audio input into segments of a certain length in order to carry out a Short-Time Fourier Transform (STFT), which takes place at the same time as the text is tokenized. The generator incorporates the payload and then passes the resulting stego representation to both the discriminator and decoder branches. Once the training optimization loop converges, an inverse Short-Time Fourier Transform (ISTFT) reconstructs the final stego audio signal in the time domain.

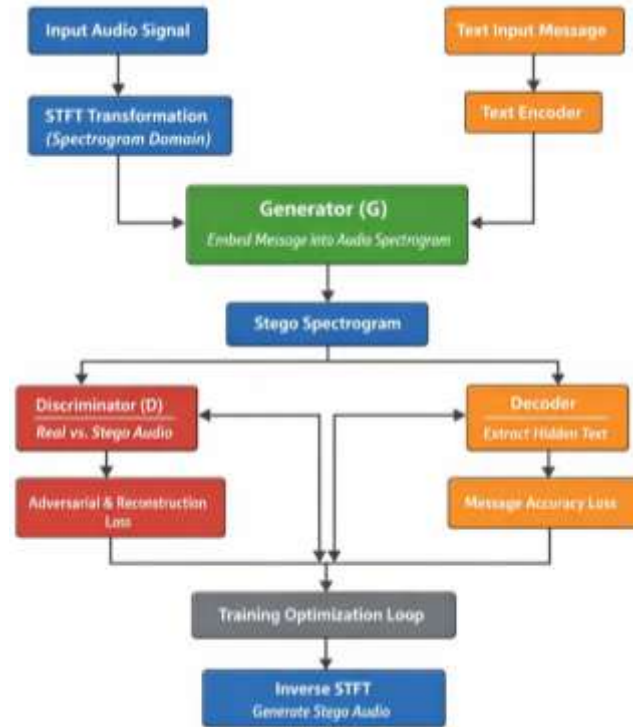


Figure 2: Flowchart of the proposed gan-based audio steganography framework.

4. Experiment

4.1. Dataset and Preprocessing

The model was trained and evaluated on a diverse audio corpus comprising:

- **The speech samples** were obtained from the LibriSpeech benchmark dataset, making a total of 500.
- **200 environmental audio samples** have been extracted from the ESC-50 dataset.

To ensure homogeneity across all training pipelines, we resampled every audio file, normalized its amplitude, and converted it to mono, 44.1 kHz, 16-bit WAV format.

4.2. Evaluation Process

The proposed modal was evaluated across imperceptibility, perceptual quality, and transmission reliability criteria using:

- The signal-to-noise ratio (SNR) measures the acoustic distortion resulting from the embedding process.
- The Perceptual Evaluation of Speech Quality (PESQ) assesses speech degradation subjectively using the standard ITU-T scales.
- The bit error rate (BER) measures the accuracy of extracting the hidden text payload under stress.
- Robustness Stress Tests: Assessed resilience against common signal perturbations, including MP3 compression, additive Gaussian noise, and low-pass filtering.

5. Results

5.1 The evaluation of imperceptibility and audio quality

To thoroughly evaluate the imperceptibility of the GAN-based audio steganography method, we computed objective acoustic fidelity metrics across the entire evaluation corpus. Incorporating the secret text messages in the spectral domain caused only slight perceptual degradation. More specifically, the framework attained the following:

- The average signal-to-noise ratio (SNR) is 28.7 dB which indicates that the original host signals and the produced stego audio files are highly mathematically similar.
- The Average Perceptual Evaluation of Speech Quality (PESQ) shows that the recovered speech tracks retain a near-transparent quality in accordance with the standard ITU-T evaluation scales.

The fact that these quantitative results obtained shows that the adversarial loss functions were effective in restricting the generator, so that audible artifacts or background white noise were not introduced during spectral manipulation.

5.2 Analysis of robustness and payload recovery

We evaluated the reliability of the extraction mechanism by exposing the stego audio files to typical transmission distortions and signal-processing attacks, and Table 1 shows the Bit Error Rate (BER) obtained under the different degradation scenarios.

Table 1: Describes Attacks types and their Bit Error Rate results.

Attack Type	Severity Parameter	Bit Error Rate (BER %)
No attack	Baseline	0.0
MP3 compression	128" kbps"	1.4
Additive noise	20dB Gaussian Noise	2.6
Filtering	3kHz Low-pass Filter	3.1

Table 1 shows that the system attains perfect recovery (BER) in ideal conditions. With regard to aggressive lossy compression (MP3), the error rate is still very low, which

demonstrates that spectral dispersion strongly protects the embedded data from frequency-bin quantization. Only minor degradations (and an increase in BER, respectively) occur due to additive Gaussian noise and low-pass filtering, thus confirming that the auxiliary decoder network offers robust error correction even when the high-frequency spectra are reduced.

6. Discussion

The experimental results show that the spectral-domain GAN architecture suggested offers improved imperceptibility and robustness when compared with traditional audio steganography methods. The fact that the average PESQ score is 4.32 proves that adversarial training allows the generator to maintain the important psychoacoustic features. Moreover, the low bit error rates obtained under lossy compression and in the case of noise attacks indicate that spreading the data out over the spectral bins enhances the resilience of retrieval.

The architecture does have some inherent limitations; the main disadvantage is the greater computational complexity that arises from optimizing the generator, discriminator, and decoder networks at the same time, leading to slower convergence. Researchers in the future will look into lightweight network distillation techniques in order to reduce the computational overhead without compromising payload capacity or retrieval accuracy.

7. Conclusion

The paper describes a new kind of generative adversarial network (GAN) approach to audio steganography that works in the spectral domain and uses Short-Time Fourier Transform (STFT) representations in order to securely hide text messages. The system achieves successful data concealment while maintaining the integrity of the host audio by employing an adversarial optimization method together with specialized techniques for reconstruction, message extraction, and ensuring spectral consistency.

Experimental evaluations show that the architecture in question attains a high degree of imperceptibility, with an average Signal-to-Noise Ratio (SNR) of 28.7 dB and a Perceptual Evaluation of Speech Quality (PESQ) score of 4.32. Moreover, when tested for robustness under different kinds of signal disturbance—such as 128kbps MP3 compression, the addition of 20dB of Gaussian noise, and 3kHz low-pass filtering—the decoder is found to preserve high retrieval accuracy together with very low bit error rates.

Even though the model shows high fidelity and resilience, future research will be directed towards optimizing the computational complexity and the time it takes for training to converge in the case of multi-network optimization, as well as at investigating lightweight neural architectures for use in real-time mobile and edge deployments.

References

- [1] D. Ye, S. Jiang and J. Huang, "Heard More Than Heard: An Audio Steganography Method Based on GAN," arXiv, 2019.
- [2] J. Yan et al., "FGAS: Fixed Decoder Network-Based Audio Steganography with Adversarial Perturbation Generation," arXiv, 2025.

- [3] "A dynamically interactable framework with dual-channel security: GAN-based speech steganography for concealed dialogues," *Knowledge-Based Systems*, vol. 330, 2025.
- [4] X. Kang, H. Zheng, J. Yang and Y.-Q. Shi, "Approaching Optimal Embedding in Audio Steganography With GAN," *ICASSP*, 2020.
- [5] J. Li, K. Wang and X. Jia, "A Coverless Audio Steganography Based on Generative Adversarial Networks," **Electronics**, vol. 12, no. 5, 2023.
- [6] Z. Al Hadad and I. H. Ali, "Survey in Image and Audio Steganography by Using Deep Learning Methods," **J. Kufa Math. Comput.**, vol. 10, no. 2, 2023.
- [7] "Steganography in Frequency Domain: Hiding Text through Audio Spectrogram," ISSN 1870-4069, 2025.
- [8] W. Zhou, J. Zhou, and S. Yang, "An Imperceptible and Robust Audio Watermarking Algorithm Based on SNGAN," *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024.
- [9] "Enhancing Audio Steganography Through CGAN-Generated Cover Audio and Adaptive LSB Embedding: A Hybrid Approach," *Malaysian Journal of Science Health & Technology (MJoSHT)*, vol. 12, no. 1, 2026.
- [10] "EVA-GAN: Enhanced Various Audio Generation via Scalable Generative Adversarial Networks for Spectral Reconstruction," *arXiv preprint arXiv:2402.00892*, 2024.