

Article

Diffusion-Based Model for Audio Steganography

Ji Xi ^{1,*}, Zhengwang Xia ¹, Weiqi Zhang ¹, Yue Xie ² and Li Zhao ³

¹ School of Computer Information Engineering, Changzhou Institute of Technology, Changzhou 213022, China; xiazhengwang@czust.edu.cn (Z.X.); zhangwq@czust.edu.cn (W.Z.)

² School of Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China; xieyue0109@njit.edu.cn

³ School of Information Science and Engineering, Southeast University, Nanjing 210096, China; zhaoli@seu.edu.cn

* Correspondence: xiji@czust.edu.cn

Abstract

Audio steganography exploits redundancies in the human auditory system to conceal secret information within cover audio, ensuring that the hidden data remains undetectable during normal listening. However, recent research shows that current audio steganography techniques are vulnerable to detection by deep learning-based steganalyzers, which analyze the high-dimensional features of stego audio for classification. While deep learning-based steganography has been extensively studied for image covers, its application to audio remains underexplored, particularly in achieving robust embedding and extraction with minimal perceptual distortion. We propose a diffusion-based audio steganography model comprising two primary modules: (i) a diffusion-based embedding module that autonomously integrates secret messages into cover audio while preserving high perceptual quality and (ii) a corresponding diffusion-based extraction module that accurately recovers the embedded data. The framework supports both pre-existing cover audio and the generation of high-quality steganographic cover audio with superior perceptual quality for message embedding. After training, the model achieves state-of-the-art performance in terms of embedding capacity and resistance to detection by deep learning steganalyzers. The experimental results demonstrate that our diffusion-based approach significantly outperforms existing methods across varying embedding rates, yielding stego audio with superior auditory quality and lower detectability.

Keywords: audio steganography; deep learning-based steganalysis; kdifffusion model



Academic Editor: Krzysztof Szczypiorski

Received: 22 August 2025

Revised: 9 October 2025

Accepted: 9 October 2025

Published: 14 October 2025

Citation: Xi, J.; Xia, Z.; Zhang, W.; Xie, Y.; Zhao, L. Diffusion-Based Model for Audio Steganography. *Electronics* **2025**, *14*, 4019. <https://doi.org/10.3390/electronics14204019>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the current era of rapid digital advancement, audio data has become an indispensable medium for daily communication, entertainment, and professional activities due to its rich information-carrying capacity and wide applicability [1]. However, during transmission and storage, audio data is vulnerable to numerous security threats, including information leakage, unauthorized eavesdropping, and malicious tampering. These threats not only compromise individual privacy and security but can also cause severe damage to corporate assets and national security [2]. Audio steganography—a covert communication technique that embeds secret information within publicly shared cover audio—offers a robust solution by enabling secure data transmission without attracting third-party detection [3]. Consequently, research on advanced audio steganography schemes holds significant theoretical and practical value, making a critical contribution to information security and societal stability.

Traditional audio steganography methods primarily utilize the temporal, frequency, and time–frequency characteristics of audio signals for information embedding, with each approach offering distinct advantages in concealment, robustness, and embedding capacity [4]. In the time domain, the least significant bit (LSB) substitution method is the most fundamental and widely used technique, embedding secret data into the least significant bits of audio samples. Recent advancements have improved traditional LSB methods; for example, Chen et al. developed an adaptive LSB algorithm that dynamically selects the embedding bit depth, enhancing concealment while achieving an error rate of only 0.5% in speech signals [5]. Frequency domain techniques include phase encoding, which conceals information by modifying phase relationships within audio segments [6]. Zhang et al. advanced this approach by increasing embedding capacity in the discrete wavelet transform (DWT) domain without compromising imperceptibility [7]. Another prominent method, echo hiding, embeds data by introducing controlled echoes into the audio signal, using parameters such as amplitude and delay to encode secret information.

Recent advances in deep learning have led to the development of novel audio steganography techniques that leverage the powerful feature extraction and modeling capabilities of neural networks, achieving significant improvements in concealment, capacity, and robustness against attacks [8]. End-to-end neural network–based steganography systems can automatically learn optimal embedding strategies through direct training on raw audio signals, enabling higher payload capacity while maintaining superior concealment [9]. Notably, attention mechanism–based approaches enhance steganographic performance by adaptively selecting optimal embedding locations [10]. From a technical implementation perspective, autoencoder architectures have become prevalent in audio steganography tasks, where an encoder embeds secret information into audio features, and a decoder reconstructs the original content. Recent innovations, such as the enhanced variational autoencoder (VAE-Stega) proposed by Yang et al. (2023), incorporate psychoacoustic models to significantly improve the imperceptibility of steganographic signals [11]. Generative adversarial networks (GANs) have made significant contributions to audio steganography. In these frameworks, generators produce encoded audio, while discriminators attempt to detect hidden information. Adversarial training continuously improves their performance. State-of-the-art GAN variants, such as the novel GAN model introduced by Chen et al., utilize the Wasserstein distance with gradient penalty to generate encoded audio with enhanced anti-detection capabilities [12]. These deep learning–based audio steganography methods not only overcome the limitations of traditional approaches in terms of capacity and concealment but also demonstrate superior resistance to attacks, providing more reliable solutions for information security applications.

In recent years, although audio steganography technology has made significant progress, it still faces numerous challenges. Traditional methods, such as LSB replacement and phase encoding, often suffer from poor concealment, low robustness, and limited embedding capacity [13]. While existing deep learning approaches have improved steganography performance to some extent, they still exhibit notable limitations. Autoencoders struggle to fully capture deep features when processing complex audio data, resulting in the insufficient concealment of embedded information. Conversely, GANs are prone to mode collapse due to their unstable training processes, which adversely affect the quality of encrypted audio generation. In contrast, the diffusion model, an emerging deep generative model, has demonstrated outstanding performance in fields such as image generation [14] and speech synthesis [15] owing to its unique ability to model diffusion processes. This model generates high-quality and diverse samples through a gradual denoising process and exhibits strong data distribution modeling capabilities. In the domain of audio steganography, the diffusion model offers significant advantages. First, its progressive generation mechanism better captures the complex distribution of

audio signals, effectively embedding secret information into deep features and substantially enhancing concealment [16]. Second, the generated encrypted audio exhibits higher quality and greater robustness against interference such as noise and compression. Additionally, through carefully designed model architectures and training strategies, larger embedding capacities can be achieved. Building on these advantages, this article proposes an innovative audio steganography scheme based on the diffusion model, aiming to overcome the limitations of existing technologies and provide more efficient and reliable solutions for information security and privacy protection. The main contributions of this paper are as follows:

- This paper proposed a novel steganographic framework based on a diffusion probability model. Compared to traditional autoencoders and GAN methods, this framework more effectively models the complex distribution of audio signals and achieves more natural information embedding.
- This paper proposed a diffusion steganography strategy that integrated a diffusion mechanism tailored to the time–frequency characteristics of audio signals. This approach effectively addresses the trade-off between capacity and robustness found in traditional methods by embedding and extracting information through a carefully designed, multi-stage diffusion process.

2. Related Work

2.1. Traditional Steganographic Approaches

The LSB steganography method, the simplest technique in audio steganography, offers advantages such as large embedding capacity and low complexity. Specifically, LSB steganography involves replacing the least significant bit of the carrier audio sample value with secret information to achieve information hiding, typically using only one least significant bit. To increase steganographic capacity, some methods utilize the two least significant bits. Gambhir et al. employed the RSA algorithm to encrypt secret information into ciphertext, then used the LSB method to embed the ciphertext into the audio signal [17]. Mishra et al. converted secret information into ASCII code and applied a genetic algorithm to identify the optimal positions for embedding the secret information within the carrier. Finally, the LSB method was used to embed the ASCII code corresponding to the secret information into the carrier [18]. Nassrulah et al. proposed an efficient audio steganography method based on the LSB, which enhances steganographic performance by balancing the hidden capacity and distortion rate of the carrier. This method adaptively determines the number of hidden bits for each audio sample based on the size of the secret information, carrier size, and signal-to-noise ratio (SNR) [19]. Although the LSB steganography method is simple and efficient, directly replacing the least significant bit with binary secret information presents significant security vulnerabilities.

Echo steganography leverages the masking properties of the human auditory system (HAS), such as real-time masking effects, to add echoes with varying delays to the carrier audio, thereby embedding secret information within it. This technique offers advantages including strong imperceptibility and straightforward implementation [1]. Oh et al. proposed an echo embedding method capable of embedding high-energy echoes without degrading the audio quality of the carrier, enhancing robustness and resistance to common signal processing modifications [20]. Erfani et al. introduced an audio watermarking technique based on echo hiding, which adds short resonances to the carrier audio to embed secret information, demonstrating good robustness against typical signal processing attacks [21]. Ghasemzadeh et al. addressed the security vulnerabilities of previous methods by improving echo hiding security through pseudo-random variation in echo parameters [22].

2.2. GAN-Based Steganography Approaches

Embedded carrier-based audio steganography using deep learning involves embedding and extracting secret information within digital audio through deep learning techniques. Based on design approaches, it can be categorized into encoder–decoder structures, automatic learning of embedding costs, and adversarial sample-based methods.

The steganography method based on an encoder–decoder structure employs a trained deep neural network to embed and extract secret information within cover audio, requiring only the training of the model itself. This approach effectively embeds and retrieves secret data in audio signals and allows for the design of improved embedding distortion costs to minimize steganographic distortion in the encrypted audio after embedding, thereby reducing anomalies caused by the embedded information [23]. Li et al. argue that image steganography models based on deep learning are unsuitable for audio steganography [24]. They utilize Gated Convolutional Neural Networks (GCNNs) [25] for encoding and decoding and propose an audio steganography model based on deep neural networks. This model incorporates the Short-Time Fourier Transform (STFT) and its inverse as differentiable layers within the network, thereby imposing critical constraints on network training.

Adversarial samples are perturbed inputs generated by adding imperceptible noise based on the gradient of the target machine learning model [26]. Carefully crafted adversarial samples can successfully deceive the model. Wu et al. argue that current CNN-based classifiers are easily fooled by adversarial samples and propose a time domain audio steganography method leveraging adversarial samples [3]. Unlike image steganography methods that heavily rely on predefined embedding costs, this approach involves different initializations of embedding costs, updating strategies during iterations, and deriving the final embedding cost from the temporary results of all iterations [27]. This method can start with a fixed or even random embedding cost and iteratively update it using adversarial attacks until improved security performance is achieved. Chen et al. proposed a method that uses adversarial samples to retrain encrypted carriers embedded by traditional methods to deceive steganalysis, addressing challenges posed by deep learning-based audio steganalysis to conventional audio steganography [5].

3. Methodology

3.1. Overall Framework

The audio steganography system consists of two primary modules: one for embedding confidential data using a diffusion-based approach and another for extracting it with the same model, as illustrated in Figure 1. Diffusion-based approaches have gained prominence as state-of-the-art methods for synthetic media creation, outperforming GANs in audio fidelity and achieving remarkable outcomes, as evidenced by research [28–30]. These systems rely on two interconnected stages: progressive noise injection and iterative reconstruction.

During the noise injection stage, structured perturbations are incrementally introduced to the original signal, converting it into a latent representation characterized by Gaussian noise distribution. The subsequent reconstruction stage employs iterative denoising to restore the corrupted data into high-quality outputs. Leveraging this mechanism, the Denoising Diffusion Probabilistic Model (DDPM) [14] revolutionized the domain by producing synthetic data with unprecedented precision, establishing a novel standard for artificial media generation.

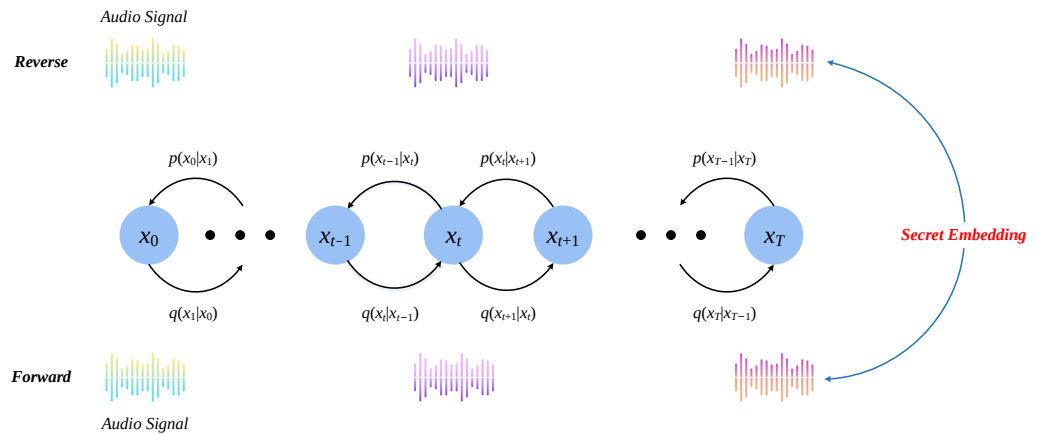


Figure 1. An overview of the proposed diffusion model for audio steganography, which comprises two primary modules: one for embedding confidential data using a diffusion-based approach and another for extracting it with the same model.

3.2. Forward Diffusion

The forward diffusion mechanism converts the original audio signal $\mathbf{x}_0$ into a latent representation by progressively injecting Gaussian noise. This process is enhanced by multi-stage diffusion steps that adaptively adjust the noise intensity across different audio segments. Unlike traditional Markovian diffusion processes, which rely on sequential transitions $q(\mathbf{x}_t | \mathbf{x}_{t-1})$, our framework directly models the marginal probability $q(\mathbf{x}_t | \mathbf{x}_0)$. This design allows for flexible control over noise intensity and facilitates non-Markovian feature fusion.

Given an original clean audio signal $\mathbf{x}_0$, the forward diffusion iteratively adds Gaussian noise over T steps. This process can be defined as follows:

$$\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon_t, \quad (1)$$

where $\epsilon_t \sim \mathcal{N}(0, 0.2)$ is a Gaussian noise vector with the same shape as $\mathbf{x}_0$.

To enhance the accuracy of noise injection, we introduce an attention mechanism that generates a weight vector to quantify the perceptual importance of audio sampling points. The process can be divided into three steps: First, the input signal $\mathbf{x}_t$ is transformed into a time–frequency spectrogram using the Short-Time Fourier Transform (STFT). Second, a psychoacoustic model assigns sensitivity scores to each spectrogram bin, prioritizing low frequencies while deprioritizing high frequencies. Finally, the sensitivity scores are mapped back to the original audio length through linear interpolation, producing attention weights. This approach integrates principles of auditory perception to emphasize perceptually critical audio segments.

During this process, the generation of $\mathbf{x}_t$ depends not exclusively on its immediate predecessor $\mathbf{x}_{t-1}$, but it also incorporates influence from the original input $\mathbf{x}_0$. The complete probabilistic formulation for this diffusion process is as below:

$$q_\sigma(\mathbf{x}_{1:T} | \mathbf{x}_0) = \prod_{t=2}^T q_\sigma(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0), \quad (2)$$

where $\mathbf{x}_0$ denotes the original clean audio input, whereas $\mathbf{x}_t$ signifies the audio corrupted by noise at iteration t . The term q_σ captures the transition probability modulated by the

parameter σ throughout the diffusion process. This parameter σ controls the intensity of stochasticity introduced during diffusion. Its definition is as follows:

$$q_{\sigma}(\mathbf{x}_T | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_T; \sqrt{\bar{\alpha}_T} \mathbf{x}_0, (1 - \bar{\alpha}_T) \mathbf{I}), \quad (3)$$

where $\mathbf{I}$ refers to the identity matrix. When $t > 1$, we can obtain the following:

$$\begin{aligned} q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \\ = N\left(\mathbf{x}_{t-1}; \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0}{\sqrt{1 - \bar{\alpha}_t}}, \sigma_t^2 \mathbf{I}\right), \end{aligned} \quad (4)$$

where $\bar{\alpha}_t$ represents a hyperparameter.

The forward diffusion process is derived using Bayesian principles:

$$q_{\sigma}(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0) = \frac{q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) q_{\sigma}(\mathbf{x}_t | \mathbf{x}_0)}{q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_0)}, \quad (5)$$

In non-Markovian diffusion frameworks, the state $\mathbf{x}_t$ is jointly determined by the initial input $\mathbf{x}_0$ and the preceding state $\mathbf{x}_{t-1}$.

After training, the neural network estimates the noise component ϵ_t through the learned function $\epsilon_{\theta}(\mathbf{x}_t, t)$. This enables the reconstruction of the original clean audio $\mathbf{x}_0$ from its noisy counterpart:

$$\mathbf{x}'_0 = \alpha(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_{\theta}), \quad (6)$$

where α is a balance parameter.

3.3. Reverse Generation

By substituting the estimated clean image $\mathbf{x}'_0$ into Equation (4), the following equation can be obtained:

$$\begin{aligned} p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t) &\approx q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0 = \mathbf{x}'_0) \\ &= N\left(\mathbf{x}_{t-1}; \left(\frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_{\theta}(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}}\right) \right. \\ &\quad \left. + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_{\theta}(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I}\right), \end{aligned} \quad (7)$$

By expanding the preceding expression, we derive the sampling equation for $\mathbf{x}_{t-1}$ given $\mathbf{x}_t$:

$$\begin{aligned} \mathbf{x}_{t-1} = & \underbrace{\sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_{\theta}(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}} \right)}_{\mathbf{x}'_0} + \\ & \underbrace{\sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \epsilon_{\theta}(\mathbf{x}_t, t)}_{\text{pointing to } \mathbf{x}_t} + \underbrace{\sigma_t \epsilon_t}_{\text{random noise}} \end{aligned} \quad (8)$$

In this scenario, σ_t is defined as η multiplied by the product of two square root terms: $\sqrt{\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}}$ and $\sqrt{1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}$, where $\eta \in [0, 1]$. The noise term ϵ_t follows a standard normal distribution $\mathcal{N}(0, 1)$, while $\epsilon_{\theta}(\mathbf{x}_t, t)$ represents its neural network-based estimate.

As specified in Equation (7), the backward generation step employs a transition probability $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ that approximates the forward diffusion process's posterior $q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$. The objective function is defined as follows:

$$\begin{aligned} -\mathbb{E}_{\mathbf{x}_0} \log p_\theta(\mathbf{x}_0) = & \mathbb{E}_{q_\sigma(\mathbf{x}_0|\mathbf{x}_T)} [\log q_\sigma(\mathbf{x}_T | \mathbf{x}_0) \\ & + \sum_{t=2}^T \log q_\sigma(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \\ & - \sum_{t=1}^T \log p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)] \end{aligned} \quad (9)$$

3.4. Training Strategy

Our approach introduces a tailored adaptation of the DDIM framework, optimized explicitly for stego audio processing. This method adheres to the original training protocol but incorporates a key modification: the preprocessing phase introduces Gaussian noise into $\mathbf{x}_0$ to ensure accurate data recovery. The hyperparameters $\tilde{\alpha}_t$ must be predefined. During each training iteration, an audio sample is randomly chosen, and synthetic noise is added to create $\mathbf{x}_0$. A diffusion time step t and additional noise ϵ are then randomly selected. The U-Net architecture predicts the noise component $\epsilon_\theta(\mathbf{x}_t, t)$ for the given step. Model parameters are optimized by minimizing the discrepancy between predicted and actual noise values, aiming to align $\epsilon_\theta(\sqrt{\tilde{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \tilde{\alpha}_t}\epsilon, t)$ with the true noise ϵ .

In this system, senders create stego audio content through a generation mechanism, while receivers employ a diffusion-based extraction process to recover embedded data with precision. This ensures true data reversibility, extending beyond mere framework-level restoration. During audio generation, as outlined in Equation (8), the random noise coefficient σ_t governs the transition from $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$. Setting $\eta = 0$ eliminates the random noise term ($\sigma_t = 0$), converting the generation into a deterministic operation. By reformulating Equation (7), the deterministic generation process can be expressed as follows:

$$\bar{\mathbf{x}}_{t-1} = \frac{\mathbf{x}_t}{\sqrt{\tilde{\alpha}_t}} + \left(\sqrt{1 - \tilde{\alpha}_{t-1}} \right) \epsilon_\theta(\mathbf{x}_t, t) \quad (10)$$

During the audio generation phase, when the starting noise audio $\mathbf{x}_T$ adheres to a standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ and is subjected to T rounds of denoising operations based on the given formula, a content-defined audio $\mathbf{x}_0$ can be produced.

By reorganizing the aforementioned Equation (10), we can obtain the transition probability $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$ for the audio generation process, as presented below:

$$\frac{\mathbf{x}_{t-1}}{\sqrt{\tilde{\alpha}_{t-1}}} = \frac{\mathbf{x}_t}{\sqrt{\tilde{\alpha}_t}} + \left(\sqrt{\frac{1 - \tilde{\alpha}_{t-1}}{\tilde{\alpha}_{t-1}}} \right) \epsilon_\theta(\mathbf{x}_t, t), \quad (11)$$

Assuming $\Delta t = 1$, Equation (11) can be rewritten as follows:

$$\frac{\mathbf{x}_{t-\Delta t}}{\sqrt{\tilde{\alpha}_{t-\Delta t}}} - \frac{\mathbf{x}_t}{\sqrt{\tilde{\alpha}_t}} = \left(\sqrt{\frac{1 - \tilde{\alpha}_t}{\tilde{\alpha}_t}} \right) \epsilon_\theta(\mathbf{x}_t, t), \quad (12)$$

Suppose $\sigma = (\sqrt{1 - \tilde{\alpha}_t} / \sqrt{\tilde{\alpha}_t})$ and $\bar{\mathbf{x}} = \mathbf{x} / \sqrt{\tilde{\alpha}_t}$. Plugging these into Equation (12) results in an ordinary differential equation (ODE):

$$d\bar{\mathbf{x}}(t) = \epsilon_\theta \left(\frac{\bar{\mathbf{x}}(t)}{\sqrt{\sigma^2 + 1}}, t \right) d\sigma(t), \quad (13)$$

Euler's method stands as a cornerstone technique for the numerical solution of ordinary differential equations (ODEs). This iterative approach approximates the solution of a differential equation by taking a sequence of minute steps. In the given equation, each $d\bar{x}(t)$ can be approximated via the Euler iteration approach. Then, using the formula $\bar{x}_{t\pm 1} = \bar{x}_t \pm d\bar{x}(t)$, we can determine the values of $\bar{x}_{t\pm 1}$. Since $x_{t\pm 1} = \sqrt{\bar{\alpha}}\bar{x}_{t\pm 1}$, we can subsequently calculate $x_{t\pm 1}$. Consequently, the audio $x_{t\pm 1}$ for any step t can be derived from the ODE, proving the reversibility of the proposed model.

3.5. Secret Data Extraction

Like the generation of stego audio, the data extraction process also makes use of accelerated sampling techniques. Given that the diffusion model is reversible, the transition probability for the accelerated diffusion process of the proposed model, $q(x_{t_{i+1}} | x_{t_i})$, can be deduced as follows:

$$x_{t_{i+1}} = \sqrt{\bar{\alpha}_{t_{i+1}}} \left(\frac{x_{t_i} - \sqrt{1 - \bar{\alpha}_{t_i}} \epsilon_{\theta}(x_{t_i}, t_i)}{\sqrt{\bar{\alpha}_{t_i}}} \right) + \sqrt{1 - \bar{\alpha}_{t_{i+1}}} \cdot \epsilon_{\theta}(x_{t_i}, t_i) \quad (14)$$

Leveraging the shared data extraction technique and the proposed model supplied by the sender, the receiver can precisely retrieve the hidden data from the stego audio x^s . In detail, the stego audio x^s , which has integer-valued signals, is initially converted into floating-point data for neural network operations. Subsequently, via the T -step noise injection process, the clean audio undergoes a transformation into a noisy audio $x^s \rightarrow x_0 \rightarrow x_1 \cdots \rightarrow x_{T-1} \rightarrow z'_s := x_T$. At this stage, the noisy audio x_T functions as the retrieved stego latent z'_s . Finally, the ultimate retrieved secrets d' can be extracted from z'_s in accordance with the extraction method:

$$\begin{aligned} \text{Coeff}_r &= \text{DCT}(z'_s) \\ d' &= \text{reshape} \left(\left\lceil \frac{\text{Sign}(\text{Coeff}_r) + 1}{2} \right\rceil \right), \end{aligned} \quad (15)$$

Here, $\text{DCT}(\cdot)$ denotes carrying out a discrete cosine transform on the input matrix. Coeff_r stands for the frequency domain coefficients acquired from the DCT of the retrieved stego latent z'_s . $\text{Sign}(\cdot)$ calculates the sign matrix of the input, which consists of values of either $+1$ or -1 . $\lceil \cdot \rceil$ operates on the input data to round it up. And $\text{reshape}(\cdot)$ restructures the input matrix into a sequence, guaranteeing that the output d' and the concealed d have identical shapes.

4. Experiments

4.1. Experimental Settings

To evaluate the effectiveness of the proposed method, this study adopted two widely recognized public datasets, TIMIT [31] and UME [32], to systematically assess its steganographic capabilities. Both datasets consist of numerous uncompressed monophonic audio recordings sampled at 16 kHz. To prepare the data for model training, all recordings underwent the following preprocessing steps.

Firstly, the min-max normalization technique was applied to normalize all audio signals, ensuring that the data could be analyzed within a standardized scale range. Secondly, the audio data were segmented into multiple small clips, each containing 16,384 time points, to facilitate subsequent model training with a batch size of 32. Finally, the Adam optimizer was employed to optimize the model, with the learning rate set to 0.0001.

4.2. Evaluation Criteria

It is worth noting that a key objective of our approach is to introduce only slight perturbations to the cover audio. That is, the cover audio should sound virtually identical to the original carrier audio. Moreover, when listening to the stego audio (generated by embedding a message into the cover audio), there should be no discernible differences from the original carrier audio.

To quantitatively evaluate the audio perception quality, we utilize well-established reference audio quality metrics, namely the subjective metric PESQ [33] and the objective metric SNR (Peak Signal-to-Noise Ratio).

- PESQ: PESQ's values span from -0.5 to 4.5 , with greater values signifying superior perceptual quality.
- SNR: The SNR represents the mean power ratio between the inherent signal and the noise.

We randomly picked 100 test audio samples from the UME dataset to serve as references, along with their corresponding cover audio for assessment. The mean PESQ score stands at 4.4235, and the SNR reaches 83.275 dB. This indicates that human ears are unable to differentiate the cover audio from the original audio, thus confirming the efficacy of our proposed approach in creating steganographic audio with excellent perceptual quality.

4.3. Comparison Methods

To demonstrate the effectiveness of our proposed approach, we carried out two experiments on the TIMIT and UME datasets. We compared the detection accuracy with LSBM [34], STC [35], two GAN-based methods [12,36], and the VAE_Stega method. Two cutting-edge deep learning-based steganalysis methods, Lin-Net [37] and Chen-Net [38], were employed to assess the undetectability of these steganography techniques.

For the TIMIT experiments, 15,000 audio clips from TIMIT were used as input for the proposed model trained on TIMIT to generate 15,000 corresponding cover audio samples. Secret bitstream messages were then embedded into these samples, resulting in 15,000 cover–stego pairs. Among them, 12,000 pairs formed the training set, and the remaining 3000 pairs were the test set. In this study, we evaluated detection performance at five different embedding rates: 0.5, 0.4, 0.3, 0.2, and 0.1 bits per sample (bps). To minimize random errors caused by dataset partitioning, the experiments were repeated 10 times, and the average of all results was taken as the final outcome.

The summarized results of this experiment are presented in Table 1. As shown in the table, the method proposed in this paper demonstrates notably low steganalysis detection rates across various datasets and under different bits-per-second (bps) conditions. These results indicate that the method is highly effective at embedding secret information into audio data. Specifically, in scenarios with low embedding rates, such as 0.1 bps, the detection accuracy of the proposed method ranges from 47.12% to 48.25%, approaching the benchmark level of random human guessing, which is 50%. This proximity to random guessing effectively underscores the method's robustness under these conditions. Further comparative analysis revealed that, compared to traditional steganography techniques such as LSBM, our proposed method demonstrated significant performance improvements, achieving increases of 12.04 percentage points at 0.1 bps and 14.18 percentage points at 0.5 bps. For example, the detection accuracies for LSBM and STC were 76.23% and 70.29%, respectively, while our method achieved 62.05%, demonstrating lower detectability. Moreover, compared with Yang et al.'s [36] GAN-based method, our proposed method attained lower detection accuracies at various embedding rates. This is because our method can generate cover audio suitable for message embedding, ensuring that the data distribution of stego audio is closer to that of the original carrier audio, making it difficult for steganalyzers to distinguish between them. For instance, when training the model on TIMIT and evaluat-

ing undetectability on UME using Chen-Net, our proposed method’s detection accuracy was 3.21% lower than Yang et al.’s [36] method at 0.5 bps. Similarly, our method exhibited better undetectability at an embedding rate of 0.1 bps. Therefore, whether compared with traditional audio steganography methods or existing GAN-based audio steganography schemes, our proposed method demonstrates excellent undetectability performance at various embedding rates.

Table 1. The detection accuracy of the proposed method with four comparison methods. In each cell, the upper value represents the detection accuracy achieved by Lin-Net [37], while the lower value corresponds to Chen-Net’s [38] accuracy. A lower detection accuracy indicates better performance in terms of undetectability.

Dataset	Steganography Method	Embedding Rates (bps)				
		0.5	0.4	0.3	0.2	0.1
TIMIT	LSBM [34]	76.13	72.37	68.79	67.22	65.06
		70.33	68.06	66.14	65.24	61.36
	STC [35]	68.34	64.86	61.35	58.75	51.32
		66.28	62.39	59.64	55.71	50.44
	Yang et al. [36]	67.37	64.64	60.17	56.36	52.07
		65.47	62.28	59.54	54.58	50.47
	Chen et al. [12]	65.18	62.37	58.79	54.43	50.04
		62.37	59.62	55.48	52.27	49.16
	VAE_Stega [11]	65.04	61.87	56.25	53.08	50.04
		62.16	58.73	55.85	52.07	49.15
	Proposed method	63.16	60.22	53.18	51.16	48.03
		60.67	56.47	54.58	51.04	48.25
UME	LSBM [34]	76.23	71.06	69.88	66.48	60.26
		72.37	68.39	65.48	62.19	58.23
	STC [35]	70.29	67.59	62.25	60.96	57.46
		66.79	64.85	61.05	57.48	54.33
	Yang et al. [36]	65.26	63.31	60.18	57.17	54.57
		66.29	64.67	60.81	57.47	52.12
	Chen et al. [12]	63.15	61.46	58.27	54.39	51.74
		61.05	58.46	56.37	53.35	49.87
	VAE_Stega [11]	62.76	60.95	56.22	52.81	50.18
		60.76	57.48	55.85	51.56	48.59
	Proposed method	62.05	59.64	54.26	51.74	48.22
		60.28	55.72	53.13	50.48	47.12

4.4. Robustness Evaluation

To verify the robustness of the proposed method against common signal processing attacks, we conducted additional experiments using the UME dataset. The stego audio generated by our diffusion model (embedding rate = 0.3 bps) was subjected to four types of attacks: MP3 compression, Gaussian noise, resampling, and low-pass filtering. We then evaluated the performance of secret information extraction, perceptual quality, and undetectability. The results are presented in Table 2.

Table 2 presents the secret information extraction performance of various audio steganography methods under Gaussian and uniform noise attacks, evaluated using the bit error rate (BER) and extraction accuracy. The proposed diffusion-based method consistently demonstrates superior robustness across all noise types and intensities, outperforming traditional techniques (LSBM, STC) as well as deep learning-based approaches (GAN-based methods, VAE Stega).

Table 2. Secret information extraction performance under attacks (BER/ Accuracy).

Attack Type	Intensity	Proposed Method	LSBM	STC	Yang et al. [36]	Chen et al. [5]	VAE_Stega
Gaussian Noise	4 dB	4.6%/95.4%	9.8%/90.2%	8.7%/91.3%	7.5%/92.5%	6.1%/93.9%	5.7%/94.3%
	8 dB	5.8%/94.2%	10.4%/89.6%	9.2%/90.8%	8.1%/91.9%	7.6%/92.4%	7.0%/93.0%
	16 dB	6.3%/93.7%	12.3%/87.7%	10.5%/89.5%	9.3%/90.7%	8.5%/91.5%	7.8%/92.2%
Uniform Noise	4 dB	4.2%/95.8%	9.1%/90.9%	8.3%/91.7%	7.2%/92.8%	6.4%/93.6%	5.3%/94.7%
	8 dB	5.5%/94.5%	9.8%/90.2%	9.6%/90.4%	7.7%/92.3%	7.2%/92.8%	6.7%/93.3%
	16 dB	6.0%/93.0%	10.6%/89.4%	9.7%/90.3%	8.8%/91.2%	8.1%/91.9%	7.4%/92.6%

Under Gaussian noise attacks, the proposed method achieves the lowest bit error rate (BER) and highest accuracy across all tested noise intensities. At 4 dB, it attains a BER of 4.6% with an accuracy of 95.4%, significantly outperforming LSBM (9.8% BER, 90.2% accuracy), STC (8.7% BER, 91.3% accuracy), and advanced methods such as Yang et al.'s [36] GAN (7.5% BER, 92.5% accuracy) and VAE Stega (5.7% BER, 94.3% accuracy). As the noise intensity increases to 8 dB and 16 dB, the proposed method maintains minimal BER growth (5.8% and 6.3% at 8 dB and 16 dB, respectively), whereas traditional methods like LSBM exhibit more severe performance degradation (10.4% and 12.3% BER at 8 dB and 16 dB).

Similar advantages are observed under uniform noise attacks. At 4 dB, the proposed method achieves a bit error rate (BER) of 4.2% (accuracy: 95.8%), outperforming LSBM (9.1% BER, 90.9% accuracy), STC (8.3% BER, 91.7% accuracy), and Chen et al.'s [5] GAN (6.4% BER, 93.6% accuracy). Even at higher noise levels (8 dB and 16 dB), its BER remains the lowest among all compared methods, confirming its robust resistance to signal distortion.

These results confirm that the proposed diffusion-based framework effectively preserves the integrity of secret information under noisy conditions, making it a reliable solution for secure audio communication in real-world environments with signal interference.

4.5. Ablation Study

In this part, we carried out ablation experiments on the key architectural variations of the proposed framework, as illustrated in Table 3. Figure 2 presents the PESQ scores achieved by various model variants after generating steganographic data using the UME dataset. Among all comparative methods, the comprehensive framework proposed in this study attained the highest PESQ score of 4.4235 for the encrypted audio. This result significantly surpasses those of the other variant methods. Specifically, the PESQ scores of other network models showed varying degrees of decline, providing strong evidence for the effectiveness of the proposed strategy. Further analysis indicates that the second and fourth schemes experienced the most pronounced performance degradation, highlighting the positive contribution of these two strategies in enhancing model performance. Overall, the experimental results reaffirm the effectiveness and superiority of the proposed method.

Table 3. The variants for the proposed diffusion method.

Index	Modified Variants
#1	Proposed framework (complete architecture)
#2	Remove the normalization operation for the input data
#3	Remove the posterior constraint
#4	Remove skip connections in the main framework
#5	Remove ReLU activation in the proposed framework

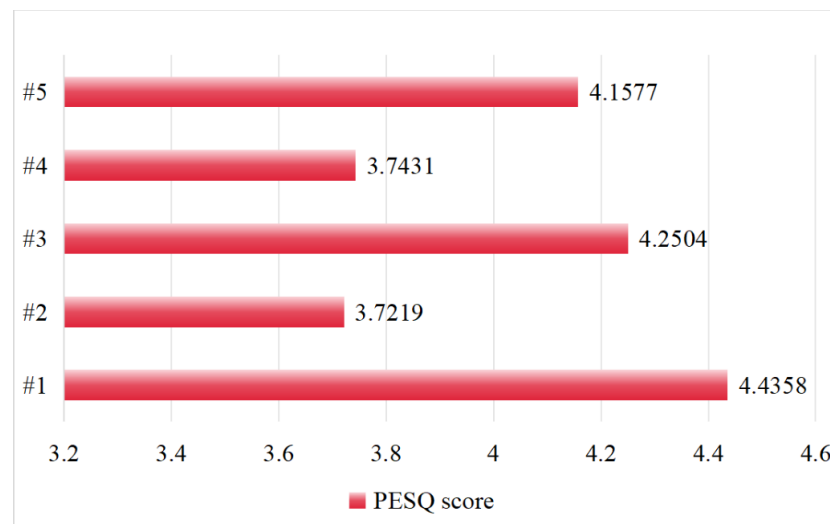


Figure 2. The statistical results of the ablation study.

5. Conclusions

In this study, we propose a diffusion-based audio steganography framework that achieves robust message embedding and extraction while maintaining high perceptual quality. The core innovation lies in two integrated modules: (i) a diffusion-driven embedding module that autonomously integrates secret data into cover audio with minimal perceptual distortion and (ii) a corresponding diffusion-based extraction module that accurately recovers embedded messages. Unlike existing methods that rely on pre-existing cover audio or suffer from detectable artifacts, our framework generates steganographic cover audio optimized for both embedding capacity and stealth. Through meticulous network design and training strategy formulation, the model achieves state-of-the-art resistance to deep learning-based steganalysis, even at high embedding rates. The experimental results confirm that our approach produces stego audio with (1) superior auditory quality (comparable to lossless compression) and (2) significantly lower detection rates across varying payloads, validating its practicality for secure audio communication.

Author Contributions: Conceptualization, J.X. and Z.X.; Funding acquisition, Z.X.; Investigation, Z.X.; Methodology, J.X. and Z.X.; Project administration, L.Z.; Software, W.Z. and Y.X.; Supervision, J.X.; Validation, J.X.; Visualization, W.Z. and Y.X.; Writing—original draft, J.X. and Z.X.; Writing—review and editing, J.X. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and Technology Plan Project of Changzhou (CJ20220151), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (23KJA520001).

Data Availability Statement: Data are available in a publicly accessible repository. The TIMIT dataset can be accessed at <https://catalog.ldc.upenn.edu/LDC93S1>, accessed on 8 October 2025, and the UME dataset is available at <https://bigballon.github.io/UME-Search/>, accessed on 8 October 2025.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bender, W.; Gruhl, D.; Morimoto, N.; Lu, A. Techniques for data hiding. *IBM Syst. J.* **1996**, *35*, 313–336. [\[CrossRef\]](#)
2. Ghasemzadeh, H.; Kayvanrad, M.H. Comprehensive review of audio steganalysis methods. *IET Signal Process.* **2018**, *12*, 673–687. [\[CrossRef\]](#)
3. Wu, J.; Chen, B.; Luo, W.; Fang, Y. Audio steganography based on iterative adversarial attacks against convolutional neural networks. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 2282–2294. [\[CrossRef\]](#)

4. Chen, K. Digital watermarking and steganography. In *Encyclopedia of Multimedia Technology and Networking*, 2nd ed.; IGI Global: Palmdale, PA, USA, 2009; pp. 402–409.
5. Chen, L.; Wang, R.; Dong, L.; Yan, D. Imperceptible adversarial audio steganography based on psychoacoustic model. *Multimed. Tools Appl.* **2023**, *82*, 26451–26463. [[CrossRef](#)]
6. Singh, L.; Singh, A.K.; Singh, P.K. Secure data hiding techniques: A survey. *Multimed. Tools Appl.* **2020**, *79*, 15901–15921.
7. Zhang, Z.; Zeng, J.; Xu, Y.; Yi, X.; Cao, Y.; Liu, C. Triple-Stage Robust Audio Steganography Framework with AAC Encoding for Lossy Social Media Channels. In Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, San Jose, CA, USA, 18–20 June 2025; pp. 131–141.
8. Wang, J.; Wang, K. A novel audio steganography based on the segmentation of the foreground and background of audio. *Comput. Electr. Eng.* **2025**, *123*, 110026. [[CrossRef](#)]
9. Subramanian, N.; Cheheb, I.; Elharrouss, O.; Al-Maadeed, S.; Bouridane, A. End-to-end image steganography using deep convolutional autoencoders. *IEEE Access* **2021**, *9*, 135585–135593. [[CrossRef](#)]
10. Peng, J.; Liao, Y.; Tang, S. Audio steganalysis using multi-scale feature fusion-based attention neural network. *IET Commun.* **2025**, *19*, e12806.
11. Yang, Z.L.; Zhang, S.Y.; Hu, Y.T.; Hu, Z.W.; Huang, Y.F. VAE-Stega: Linguistic steganography based on variational auto-encoder. *IEEE Trans. Inf. Forensics Secur.* **2020**, *16*, 880–895. [[CrossRef](#)]
12. Chen, L.; Wang, R.; Yan, D.; Wang, J. Learning to generate steganographic cover for audio steganography using GAN. *IEEE Access* **2021**, *9*, 88098–88107. [[CrossRef](#)]
13. Li, J.; Wang, K.; Jia, X. A coverless audio steganography based on generative adversarial networks. *Electronics* **2023**, *12*, 1253. [[CrossRef](#)]
14. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
15. Yang, D.; Yu, J.; Wang, H.; Wang, W.; Weng, C.; Zou, Y.; Yu, D. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2023**, *31*, 1720–1733. [[CrossRef](#)]
16. Huang, R.; Huang, J.; Yang, D.; Ren, Y.; Liu, L.; Li, M.; Ye, Z.; Liu, J.; Yin, X.; Zhao, Z. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 13916–13932.
17. Gambhir, A.; Khara, S. Integrating RSA cryptography & audio steganography. In Proceedings of the 2016 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 29–30 April 2016; pp. 481–484.
18. Mishra, A.; Johri, P.; Mishra, A. Audio steganography using ASCII code and GA. In Proceedings of the 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), Dubai, United Arab Emirates, 18–20 December 2017; pp. 646–651.
19. Nassrullah, H.A.; Flayyih, W.N.; Nasrullah, M.A. Enhancement of LSB Audio Steganography Based on Carrier and Message Characteristics. *J. Inf. Hiding Multim. Signal Process.* **2020**, *11*, 126–137.
20. Oh, H.O.; Seok, J.W.; Hong, J.W.; Youn, D.H. New echo embedding technique for robust and imperceptible audio watermarking. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (Cat. No. 01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; Volume 3, pp. 1341–1344.
21. Erfani, Y.; Siahpoush, S. Robust audio watermarking using improved TS echo hiding. *Digit. Signal Process.* **2009**, *19*, 809–814. [[CrossRef](#)]
22. Ghasemzadeh, H.; Kayvanrad, M.H. Toward a robust and secure echo steganography method based on parameters hopping. In Proceedings of the 2015 Signal Processing and Intelligent Systems Conference (SPIS), Tehran, Iran, 16–17 December 2015; pp. 143–147.
23. Fu, Z.; Wang, F.; Sun, X.M.; Wang, Y. Research on steganography of digital images based on deep learning. *Chin. J. Comput.* **2020**, *43*, 1656–1672.
24. Li, S.; Xue, M.; Zhao, B.Z.H.; Zhu, H.; Zhang, X. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Trans. Dependable Secur. Comput.* **2020**, *18*, 2088–2105. [[CrossRef](#)]
25. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 933–941.
26. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013** arXiv:1312.6199.
27. Tang, W.; Li, B.; Tan, S.; Barni, M.; Huang, J. CNN-based adversarial embedding for image steganography. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 2074–2087. [[CrossRef](#)]
28. Lemerrier, J.M.; Richter, J.; Welker, S.; Moliner, E.; Välimäki, V.; Gerkmann, T. Diffusion models for audio restoration: A review. *IEEE Signal Process. Mag.* **2025**, *41*, 72–84. [[CrossRef](#)]
29. Alexanderson, S.; Nagy, R.; Beskow, J.; Henter, G.E. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph. (TOG)* **2023**, *42*, 1–20. [[CrossRef](#)]

30. Ghosal, D.; Majumder, N.; Mehrish, A.; Poria, S. Text-to-audio generation using instruction guided latent diffusion model. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 20 August 2023; pp. 3590–3598.
31. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Tech. Rep. N* **1993**, *93*, 27403.
32. Minematsu, N.; Tomiyama, Y.; Yoshimoto, K.; Shimizu, K.; Nakagawa, S.; Dantsuji, M.; Makino, S. English Speech Database Read by Japanese Learners for CALL System Development. In Proceedings of the LREC, Las Palmas, Spain, 29–31 May 2002.
33. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (Cat. No. 01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; Volume 2, pp. 749–752.
34. Sharp, T. An implementation of key-based digital signal steganography. In Proceedings of the International Workshop on Information Hiding, Pittsburgh, PA, USA, 25–27 April 2001; pp. 13–26.
35. Filler, T.; Judas, J.; Fridrich, J. Minimizing additive distortion in steganography using syndrome-trellis codes. *IEEE Trans. Inf. Forensics Secur.* **2011**, *6*, 920–935. [[CrossRef](#)]
36. Yang, J.; Zheng, H.; Kang, X.; Shi, Y.Q. Approaching optimal embedding in audio steganography with GAN. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 2827–2831.
37. Lin, Y.; Wang, R.; Yan, D.; Dong, L.; Zhang, X. Audio steganalysis with improved convolutional neural network. In Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, Paris, France, 3–5 July 2019; pp. 210–215.
38. Chen, B.; Luo, W.; Li, H. Audio steganalysis with convolutional neural network. In Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, Philadelphia, PA, USA, 20–22 June 2017; pp. 85–90.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.