Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.036819
Article

Cover Enhancement Method for Audio Steganography Based on Universal Adversarial Perturbations with Sample Diversification

Jiangchuan Li, Peisong He*, Jiayong Liu, Jie Luo and Qiang Xia

School of Cyber Science and Engineering, Sichuan University, Chengdu, 610065, China

*Corresponding Author: Peisong He. Email: gokeyhps@scu.edu.cn
Received: 12 October 2022; Accepted: 22 February 2023

Abstract: Steganography techniques, such as audio steganography, have been widely used in covert communication. However, the deep neural network, especially the convolutional neural network (CNN), has greatly threatened the security of audio steganography. Besides, existing adversarial attacks-based countermeasures cannot provide general perturbation, and the transferability against unknown steganography detection methods is weak. This paper proposes a cover enhancement method for audio steganography based on universal adversarial perturbations with sample diversification to address these issues. Universal adversarial perturbation is constructed by iteratively optimizing adversarial perturbation, which applies adversarial attack techniques, such as Deepfool. Moreover, the sample diversification strategy is designed to improve the transferability of adversarial perturbations in black-box attack scenarios, where two types of common audio-processing operations are considered, including noise addition and moving picture experts group audio layer III (MP3) compression. Furthermore, the perturbation ensemble method is applied to further improve the attacks' transferability by integrating perturbations of different detection networks with heterogeneous architectures. Consequently, the single universal adversarial perturbation can enhance different cover audios against a CNN-based detection network. Extensive experiments have been conducted, and the results demonstrate that the average missed-detection probabilities of the proposed method are higher than those of the state-of-the-art methods by 7.3% and 16.6% for known and unknown detection networks, respectively. It verifies the efficiency and transferability of the proposed methods for the cover enhancement of audio steganography.

Keywords: Audio steganography; cover enhancement; adversarial perturbations; sample diversification

1 Introduction

Covert communication is the transmission of confidential information between the two parties, which prevents the communication signal from being discovered by malicious eavesdroppers. In the past decades, there have been various communication schemes in the field of information security,

such as information encryption, watermarking, steganography, etc. More specifically, information encryption [1,2] can be treated as a fundamental technique of security communication, where ciphertext generated by an encryption algorithm cannot be decrypted without the user's key. Next, watermarking techniques [3] are proposed to conduct the authentication of authenticity and integrity for the senders and receivers during communication. Different from them, steganography is the art of embedding secret messages into different carriers without arousing suspicion of the steganography detectors, which pays more attention to the stealthiness of the carrier for covert communication. Nowadays, steganography techniques have been widely used in covert communication. Based on the types of carriers, steganography can be divided into several categories, including steganography based on image [4], audio [5], and video [6]. Due to the widespread popularity of audio on the Internet, audio files are considered important carriers in multimedia steganography. Audio files' capacity is promising and of great advantage in practical applications.

Audio steganography and its detection method have been developed in an adversarial manner. On the one hand, the audio steganography methods were designed in the time domain at the initial stage. For instance, in the least significant bit (LSB) method [7], the least significant bits are substituted with a binary message. Later, adaptive audio steganography methods were proposed [8] to consider the cost function in the steganography code framework carefully. However, due to the time-consuming problem of adaptive audio steganography, time-domain-based audio steganography methods are still mainstream. On the other hand, the performances of audio steganography detectors have been improved significantly as a countermeasure to audio steganography by leveraging deep learning. More specifically, audio steganography detectors based on one-dimension convolutional neural networks (CNN) have achieved state-of-the-art performances to expose audio steganography, including ChenNet [9], LinNet [10], and BSNet [11]. Therefore, the security of audio steganography is under considerable threat.

In order to counter CNN-based audio steganography detectors, the adversarial attack has been used as an effective method to invalidate detection networks. These slight but intentional perturbations can fool the detection networks resulting in misclassification. An intuitive way to combine the technique of adversarial examples with steganography is to improve the security of steganography. Wu et al. [12] first proposed a novel adversarial audio steganography method focused on embedding costs, which could start from a flat or random embedding cost and then iteratively update the initial cost by exploiting the adversarial attacks. Besides, the adversarial example was applied to design an iterative adversarial post-processing model (IA-SPP) [13], which concentrated on the post-processing operations of secret audios. In this paper, the secret audio denotes the audio carrier embedded with secret messages using audio steganography methods. In the study by Ying et al. [13], adversarial perturbations were added to the third LSB layer to avoid conflict with secret messages. However, the methods mentioned above require calculating the specific adversarial perturbation for each audio sample individually, which is a time-consuming task in practical applications. Besides, they mainly focused on the attack success rate for the known (target) detection network, which was hard to transfer to unknown (non-target) detection networks.

This paper proposes a cover enhancement method for audio steganography based on universal adversarial perturbations with sample diversification to solve the limitations of previous works. The proposed method iteratively generates the universal adversarial perturbation (UAP) based on adversarial example techniques, such as Deepfool [14]. In general, UAP takes a specific attack success rate as the optimization objective, which only requires constructing a single universal adversarial perturbation to enhance different cover audios. Furthermore, a sample diversification strategy and a perturbation

ensemble method are proposed to improve the transferability of adversarial perturbations in black-box attacks before and after the perturbation construction, respectively. The main contributions of this work are summarized as follows:

1. This paper proposes a cover enhancement method for audio steganography based on universal adversarial perturbations with sample diversification. To the authors' best knowledge, it is the first work to conduct audio cover enhancement with adversarial attacks. It iteratively constructs the universal adversarial perturbation based on adversarial example techniques, such as Deepfool, and applies a specific attack success rate as the optimization objective.

2. This paper designs a sample diversification strategy (SDS) inspired by the input transformation technique to produce subtle changes in the input patterns by using common audio processing operations, such as noise addition and moving picture experts group audio layer III (MP3) compression, to improve the transferability of adversarial perturbations in black-box attack scenarios.

3. This paper proposes a perturbation ensemble method (PEM) to integrate the universal adversarial perturbations constructed by detection networks of audio steganography with heterogeneous architectures, thus improving attack transferability.

4. This paper conducts extensive experiments to evaluate the security performance against known and unknown detection networks. Experimental results verified that audio steganography methods equipped with the proposed scheme could achieve better security than their original versions, especially for black-box attacks.

The rest of the paper is organized as follows: Section 2 briefly introduces the related works. Section 3 describes the proposed generation scheme. Section 4 evaluates the proposed method through extensive experiments. Section 5 draws the conclusions.

2 Related Works

In this section, audio steganography, the detection method of audio steganography, and cover enhancement with adversarial attacks are introduced separately.

2.1 Audio Steganography

Digital audio steganography is an important covert communication method, where most techniques were proposed in the time domain. For instance, the least significant bit (LSB) substitution was first applied for audio steganography by Sridevi et al. [7]. After that, a new encoding LSB-based technique [15] was designed to reduce the embedding distortion with an increased capacity. More advanced, Ahmed et al. [16] increased the embedding layer depth from the fourth LSB layer to the eighth LSB layer without noticeable perceptive artifacts. Next, Kumar et al. [17] designed a strategy that selected sample bits according to the Fibonacci numbers and then modified the corresponding LSB to hide data. Kar et al. [18] presented a scheme for embedding secret message bits in chaotical distribution based on selected thresholds. But the security performances against detection networks of audio steganography must improve. Later, inspired by the adaptive image steganography methods, Luo et al. [19] proposed adaptive audio steganography based on advanced audio coding (AAC), which can obtain relatively better security performances. However, it is much more time-consuming than time-domain-based methods. Then, Chen et al. [20] designed a distortion cost through the "large-amplitude-first" rule. Due to the time efficiency in practical applications, this paper focuses on time-domain-based audio steganography in this work.

2.2 Detection Method of Audio Steganography

The detection methods of audio steganography aim to expose the existence of secret messages, including hand-crafted feature-based detectors and deep learning-based detectors, which significantly challenge audio steganography. For detection methods based on hand-crafted features, Liu et al. [21] presented an approach based on the Mel-cepstrum coefficients derived from the second-order derivative. Then, Luo et al. [22] designed a detection feature set derived from both time and frequency domains with an ensemble classifier [23]. Recently, an end-to-end detection framework of audio steganography based on 1D-CNN was proposed to prove the advantages of deep learning-based detectors over hand-crafted feature-based detectors [9]. Lin et al. [10] designed an improved CNN-based method with a high-pass filter (HPF) and truncated linear unit (TLU). Zhang et al. [24] considered the deeper structure and complicated convolutional modules to construct a deep residual convolutional network. Lee et al. [11] proposed a CNN with bit-plane separation to directly expose the modified bits in audio samples, which showed good detection performance for low embedding rates.

2.3 Cover Enhancement with Adversarial Attacks

Cover enhancement is one of the security enhancement methods for steganography, especially by leveraging adversarial attacks. Inspired by the success of adversarial attacks to fool neural networks in computer vision [25], researchers have applied adversarial attacks to generate adversarial covers. Compared with the original covers, the secret information can be embedded in the adversarial covers with better security performance. Several related works have been proposed in image steganography. More specifically, Zhang et al. [26] generated adversarial perturbations using the gradients of secret images based on the fast gradient sign method (FGSM) [27], which resulted in misclassifying adversarial secret images as cover images. Zhou et al. [28] proposed a framework based on the generative adversarial network (GAN) [29] to obtain adversarial perturbations that enhance covers and designed a loss function to improve security. Qin et al. [30] proposed a sparse cover enhancement method, which effectively compressed the distortions through sparse ± 1 adversarial perturbation and a retrying scheme. However, there is a lack of cover enhancement for audio steganography. Chen et al. [31] proposed a cover audio generation method based on GAN directly instead of enhancing existing covers, which is the most related work to enhance the security of audio steganography. However, it causes more distortion and is more time-consuming to generate covers.

3 Methodology

As mentioned above, cover enhancement methods have several advantages, including 1) less distortion introduced to the properties of original covers compared with cover generation methods; 2) the more promising time efficiency. For audio steganography, to deal with the limitations mentioned above, this paper proposes a cover enhancement method based on universal adversarial perturbations with sample diversification.

To the authors' best knowledge, it is the first work of enhancement method for cover audios. In this section, the proposed cover enhancement method based on universal adversarial perturbation (UAP) is introduced in detail. Fig. 1 shows the proposed framework of cover enhancement for audio steganography. At first, a set of clean cover audio is presented, which is enhanced by the sample diversification strategy (SDS) to generate the diversified cover audio. Then, a set of diversified secret audio corresponding with diversified cover audio is presented, which is generated by embedding secret messages into diversified covers with the audio steganography method (such as LSBM [32]). The generation of the audio UAP needs to use all diversified secret audios and gradually build

the UAP. This kind of perturbation is called "universal" since a single perturbation can process different audio samples. Then, a perturbation ensemble method (PEM) is applied to obtain the final ensemble perturbation based on UAPs generated by different detection networks with heterogeneous architectures to improve the transferability in black-box attacks. The final ensemble perturbations are added to the samples in the diversified cover audios instead of the diversified secret audios to ensure the successful extraction of secret messages. Therefore, the adversarial cover audios are constructed by adding the final ensembled perturbation to the diversified cover audios, and then the adversarial secret audios are obtained by embedding secret messages into adversarial cover audios. The goal of this framework is to make adversarial secret audios able to fool the steganography detection networks, which target to achieve better security performance.

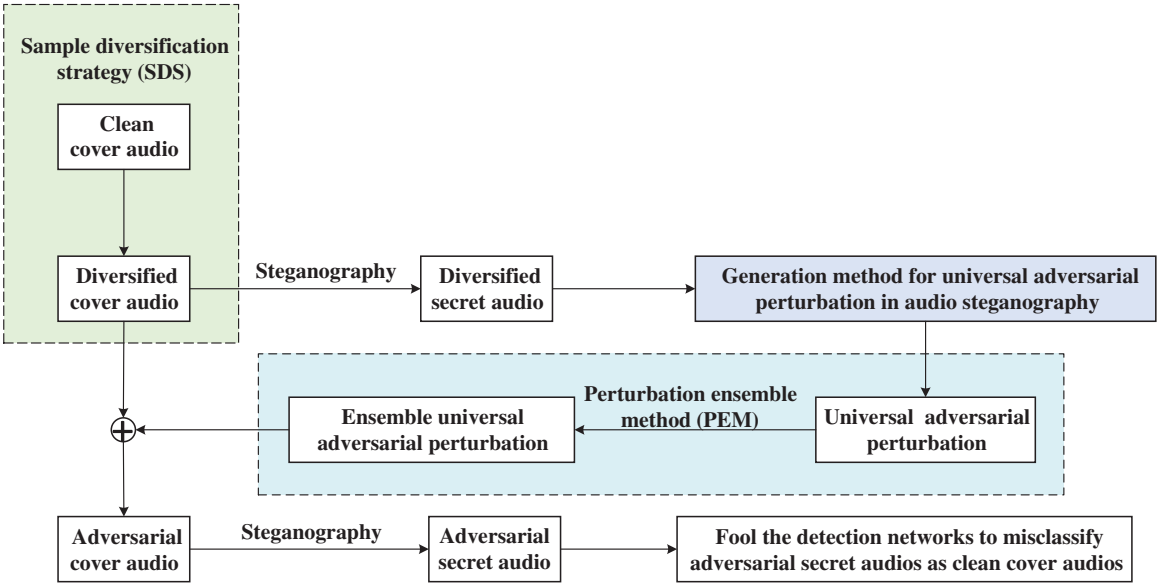

Figure 1: The proposed framework of the cover enhancement method for audio steganography

3.1 Sample Diversification Strategy

In the field of adversarial attacks, researchers have found that preprocessing operations of training samples are necessary to improve adversarial attacks' transferability. For example, Xie et al. [33] applied common image transformations to process input samples to alleviate the overfitting phenomenon and achieve higher success attack rates in black-box attacking scenarios. Inspired by these ideas, this paper designs a sample diversification strategy (SDS) for audio cover enhancement to improve the transferability in black-box attacks, which only makes subtle changes to reduce overfitting before constructing the UAP. Besides, due to leveraging common audio processing operations, SDS does not cause malicious distortions. The detailed procedures are presented as follows:

Deep learning-based detectors of audio steganography can be treated as binary classifiers. In the ideal condition, cover audios can be distinguished from secret audios by the classification boundary in the feature space. Assume the authors have a set of cover audio samples. Then, clean cover audios are enhanced to generate diversified cover audios using SDS. Subsequently, secret messages are embedded into samples of diversified cover audios to obtain diversified secret audios with a specific steganography method. Generally, for adversarial attack techniques, iterative attacks greedily

perturb the inputs in the direction of the sign of the gradient at each iteration. However, it is easy to fall into the poor local maxima and overfits the target network. This kind of overfitted adversarial perturbation is hard to transfer to unseen models (detection networks) in black-box attack cases. Training samples (clean cover audios) of generating adversarial perturbations are preprocessed, and then obtain diversified cover audios to deal with this issue. This procedure can be presented as follows:

$$\mathbf{c}_i = T(\mathbf{x}_i) \tag{1}$$

where T donates the transformation function of SDS, $\mathbf{x}_i$ represents the ith clean cover audio, and $\mathbf{c}_i$ represents the ith diversified cover audio. Therefore, the proper selection of transformation functions is significant for SDS.

In practical applications, there are many types of common audio-processing operations that can be applied to enhance audio samples, such as noise addition, reverberation addition, waveform shifting, waveform stretching, pitch shifting, etc. However, for the detection task of audio steganography, abnormal traces of high-frequent components are the most important clues of different detection networks instead of sound intensity and loudness. Therefore, to achieve better transferability against unseen detection networks, preprocessing operations that conduct slight distortions on audio samples, especially for high-frequent components, should be considered. In this work, noise addition and MP3 compression are selected for SDS based on the purpose mentioned above.

Noise addition. Gaussian white noise is selected to conduct the noise addition. Specifically, Gaussian white noise is a random signal whose instantaneous value obeys a Gaussian distribution and whose power spectral density is uniformly distributed. The components in a Gaussian white noise are statistically independent. In this paper, the first transformation function is to add the Gaussian white noise (with a mean of 0, a standard deviation of 1, and the magnitude controlled by noise factor) into the cover audio. The noise factor should be set as a small value related to the magnitude of audio samples. For each sample, Gaussian white noise is multiplied by the noise factor and then added to clean cover audios to obtain diversified cover audios. Fig. 2 shows an example, including the clean cover before noise addition, the diversified cover after noise addition, and the difference between them.

MP3 compression. MP3 compression is an entirely different way compared with noise addition to diversifying input samples. For the convenience and efficiency of storage and transmission, raw audio signals in lossless format (such as WAV) are compressed into the lossy format in most cases, such as MP3. More specifically, MP3 is one of the widely-used compression techniques. It contains several steps, including framing and sub-band filter, psycho-acoustic model, modified discrete cosine transform, quantization, Huffman encoding, and bitstream formatting.

For an audio file, its bit rate depends on the sampling rate. For instance, if the sampling rate and bit depth of an audio file are 16 kHz and 16-bit, its bit rate is 256 kbps. The compression ratio is defined to control the strength of lossy compression, which denotes the ratio of the original bit rate to the compressed bit rate. It should be noted that the degradation of the original audio signal caused by MP3 compression is related to the compressed rate and the properties of local contents. The increment of training samples' diversity aims to improve adversarial attacks' transferability against unseen detection networks, which pay more attention to high-frequency components. Therefore, the compression ratio of constructing diversified cover audios is set to 1.0, 2.0, and 3.0, respectively. Fig. 3 shows an example, including the clean cover before MP3 compression, the diversified cover after MP3 compression, and the difference between them.

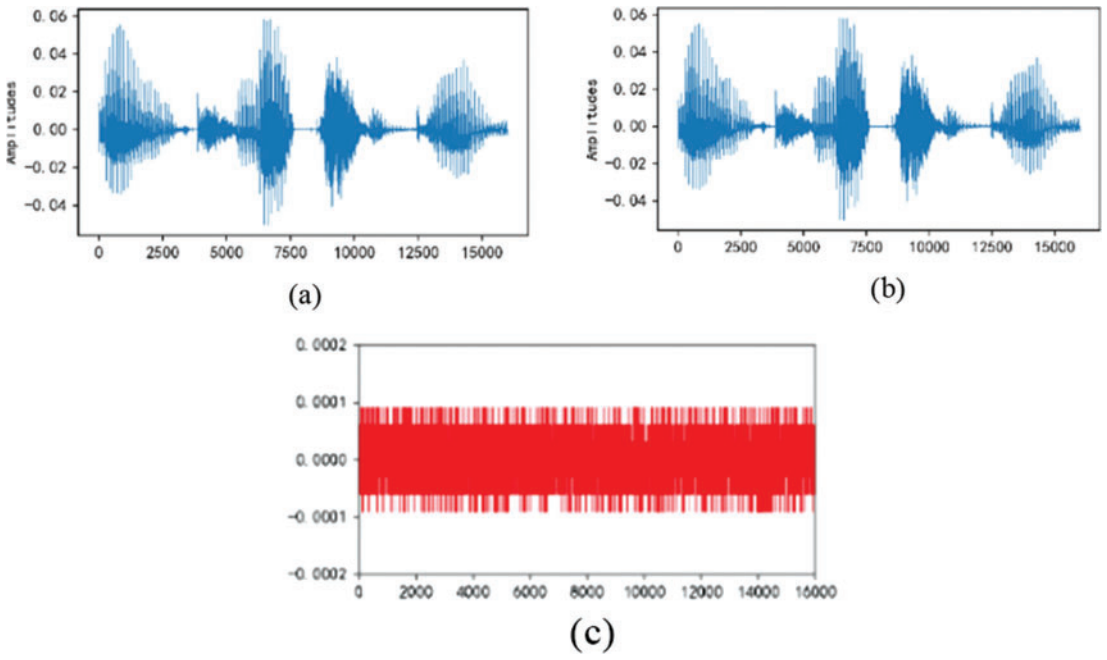

Figure 2: Clean cover before noise addition (a), diversified cover after noise addition (b), and the difference between them (c)

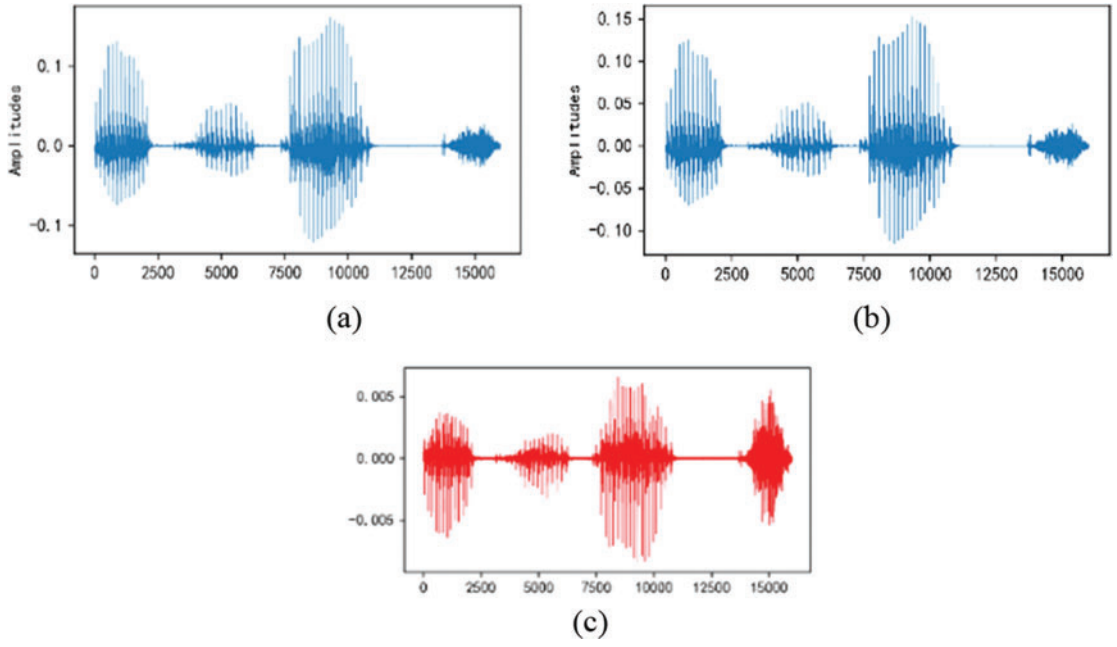

Figure 3: Clean cover before MP3 compression (a), diversified cover after MP3 compression (b), and the difference between them (c)

3.2 Generation Method for Universal Adversarial Perturbation of Audio Steganography

Universal adversarial perturbation is an effective type of adversarial attack, which can fool deep neural networks into making misclassification for different input samples with a single perturbation. Moosavi-Dezfooli et al. [34] first proposed a greedy algorithm to obtain such perturbations for image classification tasks by aggregating perturbation vectors. Abdoli et al. [35] extended this idea and designed a novel penalty formulation to find universal adversarial perturbations for audio classification systems. Inspired by these ideas, a universal adversarial perturbation for audio steganography is attempted to develop. In this subsection, the generation method for universal adversarial perturbation of audio steganography is presented.

The critical point is to obtain universal adversarial perturbation (UAP), which can fool the target detection network on diversified secret audios. The diversified secret audio is obtained by embedding the secret message into the corresponding cover audio by a specific audio steganography method, such as LSBM [32]. The problem of calculating the UAP can be defined as making the target detection network misclassify secret audios. The perturbation is called "universal" because a single perturbation can enhance different cover audios. Therefore, two constraints are considered to construct the UAP:

$$\|\mathbf{v}\|_\infty \leq e \tag{2}$$

where $\mathbf{v}$ denotes the UAP and e denotes the threshold for the maximum norm of the perturbation to control its magnitude. The second constrain is formulated as follows:

$$P\{N(\mathbf{s}) = c_0\} \geq d \tag{3}$$

where N denotes the detection network, $\mathbf{s}$ denotes diversified secret audio, and d denotes the desired success rate of attacks. P denotes the probability of misclassifying the diversified secret audio as a cover sample. The detailed procedure for calculating the UAP is presented as follows and listed in Algorithm 1. Here, c_0 denotes the label of cover audio, which is set as 0, while c_1 denotes the label of secret audio, which is set as 1.

The UAP is initialized as a zero vector, whose size is consistent with the secret audio, and then updated iteratively.

During one iteration, for each diversified secret audio, the perturbed vector obtained by adding the UAP into the diversified secret audio is fed into the target detection network. On the one hand, if the UAP cannot fool the target detection network to misclassify the secret audio as a cover sample, a new perturbation to update the UAP is calculated according to Eq. (4):

$$\begin{aligned} &\mathbf{v}^* \leftarrow \underset{\mathbf{v}_e}{\operatorname{argmin}} \|\mathbf{v}_e\|_2 \\ &s.t. N(\mathbf{s} + \mathbf{v} + \mathbf{v}_e) = c_0 \end{aligned} \tag{4}$$

where $\mathbf{v}_e$ denotes the updated UAP, and $\mathbf{v}*$ denotes the minimum UAP.

Existing generation methods of adversarial examples can obtain the updated UAP. Here, this paper selects DeepFool [14] to obtain the minimum UAP, which considers the magnitude of the perturbation and attack performance. More details can be found in [14]. Then, the obtained minimum perturbation is added to the current UAP to update its value, which can push the perturbed vector toward the target side of the decision boundary of the detection network in the feature space.

On the other hand, if the current perturbation can fool the detection network successfully, the value of m (the number of misclassified samples) will be increased by one. Besides, the final minimum UAP should be adjusted based on the first constraint about the magnitude, which can be formulated

as the projection on the maximum norm ball with the magnitude constraint and centered at 0. It is shown in Eq. (5):

$$P(\mathbf{v}') = \underset{\mathbf{r}}{\operatorname{argmin}} \|\mathbf{v}' - \mathbf{r}\|_2$$
$$s.t. \|\mathbf{r}\|_\infty \leq e \text{ and } \mathbf{v}' = \mathbf{v} + \mathbf{v}^* \tag{5}$$

The optimization problem Eq. (5) is solved by comparing the absolute value of each element in the final optimal perturbation ($\mathbf{v}'$) with the magnitude constraint. If the absolute value is smaller than the magnitude constraint, the element in the final optimal perturbation keeps unchanged. Otherwise, the element is updated by the magnitude constraint with its current sign. The result of Eq. (5) is the final perturbation for diversified secret audios and is denoted as UAP ($\mathbf{v}$) in this iterative optimization process.

At the end of an iteration, for the final optimal perturbation, the attack success rate (ASR) is calculated where ASR=m/t. m denotes the number of misclassified samples, while t denotes the number of all samples during one iteration. When the ASR on perturbed secret audios exceeds the desired success rate of attacks, the iterative optimization of the UAP is finished, which can be shown as follows:

$$\text{ASR}(\mathbf{v}) \geq d \tag{6}$$

Besides, when the iteration number u reaches the maximum value L, the algorithm stops and outputs the UAP. Otherwise, it starts a new iteration.

Algorithm 1 Generation method for universal adversarial perturbation of audio steganography

Input: diversified secret audios ($\mathbf{s}$) (t is the total number of diversified secret audios), the target detection network (N), the allowed magnitude constraint (e), the desired success rate of attacks (d), and the maximum number of iterations (L).
Output: universal adversarial perturbation ($\mathbf{v}$).

Initialize $\mathbf{v} = 0$, $m = 0$, $u = 0$
while ASR ($\mathbf{v}$) $\leq d$ and $u \leq L$ **do**
  for each $\mathbf{s}$ of the diversified secret audios **do**
    if $N(\mathbf{s} + \mathbf{v}) \neq c_0$ **then**
      // Compute minimal perturbation that leads to misclassification
      $\mathbf{v}^* \leftarrow \text{argmin}_{\mathbf{v}_e} \|\mathbf{v}_e\|_2 \, s.t. N(\mathbf{s} + \mathbf{v} + \mathbf{v}_e) = c_0$
      // Update the perturbation $\mathbf{v}$
      $\mathbf{v} \leftarrow P(\mathbf{v} + \mathbf{v}^*)$
    else $m := m + 1$
    end if
  end for
  ASR $\leftarrow m/t$
  $u := u + 1$
end while

3.3 Perturbation Ensemble Method

It has been studied that a strong adversarial example can achieve a high success rate in the white-box attack scenario, where detection networks are known for the proposed task. However, the attack performance usually declines when detection networks are unknown, which infers poor transferability.

In this framework, the perturbation ensemble method (PEM) is designed to further improve the transferability of black-box attacks by integrating UAPs generated by different detection networks with heterogeneous architectures. More specifically, the framework of the perturbation ensemble method is shown in Fig. 4. PEM can fuse the features of different UAPs to overcome the shortage of transferability for a single UAP. The construction of an ensembled UAP ($\mathbf{v}_u$) can be formulated as follows:

$$\mathbf{v}_u = \sum_{i=1}^{K} a_i * \mathbf{v}_i \tag{7}$$

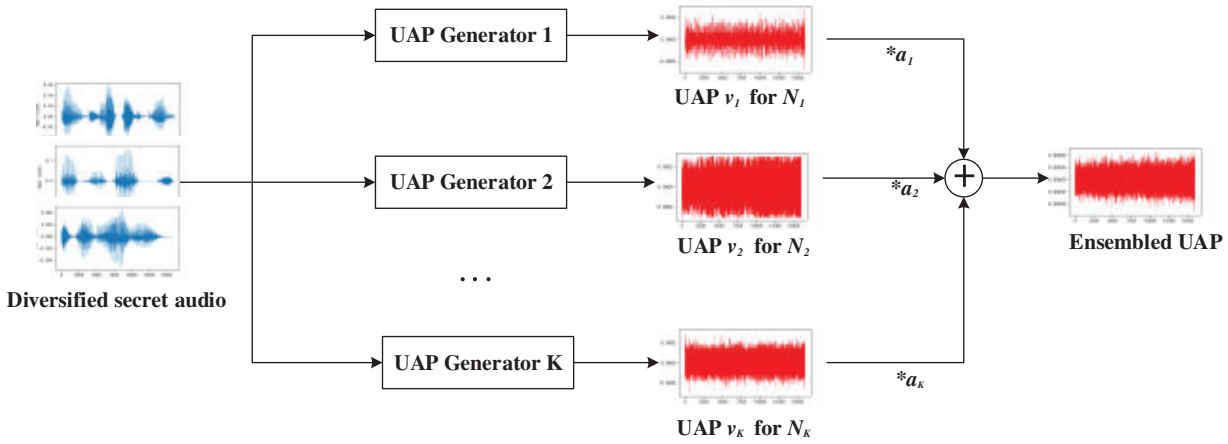

Figure 4: The framework of the perturbation ensemble method

where a_i denotes the weight of $\mathbf{v}_i$ and $\mathbf{v}_i$ is the UAP generated for the ith detection network. K denotes the total number of detection networks. In this work, three state-of-the-art audio detection networks with heterogenous architectures, including ChenNet, LinNet, and BSNet, are utilized to generate the ensembled UAP. These three different UAPs generated based on ChenNet, LinNet, and BSNet are used to obtain four types of ensembled UAPs, including ChenNet + LinNet, LinNet + BSNet, ChenNet + BSNet, and ChenNet + LinNet + BSNet. More specifically, ChenNet applied convolutional layers with the fixed parameters of high-pass filters to suppress audio content and adaptively captured the slight modification caused by embedding secret messages. Besides, a mixture of the convolutional layer and max pooling layer was used to perform subsampling to achieve good representations and prevent overfitting. LinNet designed a high-pass filter (HPF) layer to expose steganographic artifacts and a Truncated Linear Unit (TLU) as the activation function. BSNet applied bit-plane separation, weight-standardized convolution, and channel attention to develop a CNN with better detection performance.

4 Experiment Results and Analysis

In this section, extensive experiments are conducted to verify the feasibility and effectiveness of the proposed method.

4.1 Dataset

In the experiment, the audio dataset is constructed based on the TIMIT speech corpus [36], which consists of 6,300 raw audio files stored in the WAV format with 16 kHz and 16-bit quantization.

The preprocessing procedure of the dataset is the same as that of BSNet. First, audio samples whose duration is shorter than 1 sec are discarded. Consequently, 6,298 speech audios are valid. Then, for each speech audio, ten audio chunks with a one-second duration are randomly extracted. Finally, 62,980 audio samples are obtained. The audio samples are divided into two subsets for training and testing with a ratio of 3:2. The training subset is used to generate the UAP, while the testing subset is used to evaluate the security performance.

4.2 Evaluation Measure

In this work, the typical metric of cover enhancement is adopted to evaluate the performance, namely missed-detection probability (P_m). It has been mentioned that cover enhancement is used to fool the detection network and improve the security of steganography, where the detection of audio steganography can be treated as a binary classification task. Specifically, cover audios and secret audios are negative and positive samples, respectively. Therefore, P_m refers to the proportion of secret audios misclassified as cover audios overall secret audios, which is computed as:

$$P_m = \frac{F}{S} \tag{8}$$

where F denotes the number of secret audios misclassified as cover audios while S denotes the total number of secret audios. When the value of P_m is larger, the detection network is more likely to misclassify secret audios as cover audios, which indicates the better security of steganography methods. Please note that this paper didn't consider the detection error rate of cover audios (negative samples) since cover enhancement is not conducted on cover audios.

The Signal-to-Noise Ratio (SNR) is applied as the metric to evaluate the noise level concerning the original signal. It can also measure the level of distortion of the signal after adding the UAP. This metric is measured in dB and computed as follows:

$$\mathrm{SNR}\,(\mathbf{x}, \mathbf{y}) = 20 \log_{10} \frac{P_w\,(\mathbf{x})}{P_w\,(\mathbf{y})} \tag{9}$$

where $\mathbf{x}$ denotes the audio signal and $\mathbf{y}$ denotes the noise signal. P_w is the function to calculate the power of a signal, which is defined as:

$$P_w\,(\mathbf{y}) = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \mathbf{y}\,(n)^2} \tag{10}$$

where $\mathbf{y}(n)$ donates the nth component of a vector. A higher value of SNR indicates a slighter distortion caused by the perturbation.

4.3 Experimental Settings

In experiments, three state-of-the-art audio detection networks, including ChenNet, LinNet, and BSNet, are utilized to evaluate security performance. Besides, two typical audio steganography methods are considered, including LSB matching (LSBM) [32] and LSB replacement (LSBR) [37]. These LSB-based algorithms leveraged a scattered mode to embed secret messages in random order with a seed value. Besides, two embedding rates are considered, including 1.0 bps (bit per sample) and 0.5 bps for steganography algorithms.

To the authors' best knowledge, the proposed method is the first work to conduct cover enhancement for audio steganography. Therefore, in the comparison experiment, this paper constructed a cover

enhancement method based on FGSM [27], which extended a cover enhancement method of image steganography [30] from two-dimensional data (grayscale image) to one-dimensional data (audio). Besides, the advantage of the proposed method is also evaluated by comparing the enhanced covers using the proposed UAP with their original version.

The magnitude constraint is set to 10 for target and black-box attack scenarios according to the range of elements in audio samples with 16-bit quantization. The maximum number of iterations is only set as 10, considering the balance of time efficiency and attack performance. The fusion weight is set to 1.0 for all networks in the perturbation ensemble method.

4.4 Performance Evaluation of Universal Adversarial Perturbation

In this section, the performances of UAP (without SDS and PEM) are evaluated for attacking known and unknown detection networks, respectively. Then, the SNR and the time efficiency of the proposed method are also evaluated.

4.4.1 Attacking Known Detection Networks

In the proposed method, the generation method of UAP for audio steganography is one of the significant contributions. This paper compares UAP with the modified version of cover enhancement based on typical FGSM [27] to deal with one-dimensional data (audio) to prove the security performances of UAP in cover enhancement. Table 1 shows the security performances of cover enhancement methods for different steganography methods against target detection networks. The better results are in bold.

Table 1: Security performances (P_m) by applying different cover enhancement methods

Cover enhancement	Steganography	Detection networks	0.5 bps	1.0 bps
Baseline	LSBM	ChenNet	0.0572	0.0148
		LinNet	0.2085	0.1239
		BSNet	0.0951	0.0472
	LSBR	ChenNet	0.0294	0.0113
		LinNet	0.2435	0.0921
		BSNet	0.0759	0.0443
FGSM [27]	LSBM	ChenNet	0.8024	0.8214
		LinNet	**0.8550**	0.7630
		BSNet	**0.7761**	0.6975
	LSBR	ChenNet	0.7763	0.8136
		LinNet	0.8027	0.8176
		BSNet	0.6249	0.6417

(Continued)

Table 1: Continued

Cover enhancement	Steganography	Detection networks	0.5 bps	1.0 bps
UAP (Proposed)	LSBM	ChenNet	**0.8569**	**0.8906**
		LinNet	0.8209	**0.8220**
		BSNet	0.7449	**0.7618**
	LSBR	ChenNet	**0.8364**	**0.8740**
		LinNet	**0.8726**	**0.8785**
		BSNet	**0.6731**	**0.6987**

In the column of cover enhancement, "Baseline" refers to embedding secret messages without cover enhancement. "FGSM" refers to the FGSM-based cover enhancement method. "UAP" refers to the proposed cover enhancement method only applying UAP. Here, the magnitude of perturbation generated by FGSM is 10, which is identical to that of UAP. Since cover enhancement is not conducted on cover audios considering practical conditions, this paper didn't present the detection error rate of cover audios. As shown in Table 1, the distinct improvement of P_m demonstrates the superiority of UAP, even when the proposed SDS and PEM are not applied in this experiment. According to the results in Table 1, three observations can be concluded:

- The P_m of UAP is larger than FGSM and Baseline in almost all cases, which indicates better security. The security performance of "Baseline" (without cover enhancement) is much worse than two cover enhancement methods, including FGSM and UAP, for all detection networks. This result infers that it is significant to conduct cover enhancement for audio steganography. The proposed method can achieve a distinct improvement compared with FGSM. For example, the P_m of this method is larger than that of FGSM by about 7% for LSBM with the 1.0 bps embedding rate against ChenNet. It demonstrates the iterative generation strategy of UAP is more efficient in obtaining adversarial perturbation for audio steganography.
- The improvements achieved by applying cover enhancement methods for ChenNet in different cases are greater than those of LinNet and BSNet. It may be due to the fact that the network structure of ChenNet is relatively simple and more likely to be attacked. For ChenNet, the improvements of two enhancement methods are nearly 80% more than Baseline in all cases, while for two other detection networks, the average improvements are about 66%.
- For all given steganography methods, the improvements of P_m achieved by applying cover enhancement methods increase when the embedding rates are enlarged. It is because higher embedding rates lead to modifying more data points in audio samples, which become easier to be identified by detection networks.

4.4.2 Attacking Unknown Detection Networks

In existing methods, the adversarial perturbation is generated based on one target detection network. Considering practical applications, the steganography detectors may apply different networks for analysis. Therefore, it is important to evaluate the attack transferability against unknown detection networks. In this experiment, the typical method LSBM with 1.0 bps is considered. Table 2 presents the security performances of cover enhancement methods for attacking unknown detection networks. The better results are in bold. Networks used to generate the UAP (N_a) are listed in the second column, and networks used to evaluate the security performance (N_s) are listed in the first row. In Table 2, the results in black-box attack scenarios are presented in the off-diagonal positions for each cover enhancement. The following observations are obtained, comparing the results in Tables 1 and 2:

- The P_m of UAP is better than those of FGSM and Baseline in most cases, which infers better transferability of attacks. The security performances of Baseline are slightly better than the two cover enhancement methods based on adversarial attacks except when N_a is ChenNet and N_s is LinNet. This phenomenon may be due to the fact that the network architecture of LinNet can be regarded as an advanced version based on ChenNet, which may suffer from the overfitting problem of adversarial perturbations to cause negative transferability. For the proposed UAP, the most significant improvement of P_m is achieved when N_a is LinNet and N_s is BSNet, higher than FGSM by about 9%.
- Although UAP can improve transferability in most situations, the improvement of transferability for different networks is dissimilar. For example, when N_a is LinNet and BSNet, and N_s is ChenNet, the P_m is increased by 0.4232 and 0.3641, respectively. This result demonstrates that it is easier to attack LinNet. It may be because the structure of HPF in LinNet contributes to good performances in preventing overfitting, whose UAP has the best transferability among these three detection networks.
- The P_m values for unknown networks are worse than those for known networks, as shown in Table 1. It is reasonable that the target UAP owns some typical features from target networks, where the prior knowledge of these target networks can improve the security effectively by adding adversarial perturbation.

Table 2: Security performances (P_m) by applying different cover enhancement methods against unknown detection networks

Cover enhancement	N_a	N_s		
		ChenNet	LinNet	BSNet
Baseline	ChenNet	–	**0.1239**	0.0472
	LinNet	0.0148	–	0.0472
	BSNet	0.0148	0.1239	–
FGSM [27]	ChenNet	–	0.1054	0.1293
	LinNet	0.3544	–	0.5221
	BSNet	0.2971	0.2439	–
UAP (Proposed)	ChenNet	–	0.1065	**0.1722**
	LinNet	**0.4380**	–	**0.6078**
	BSNet	**0.3789**	**0.2660**	–

4.4.3 Performance Evaluation of SNR

In this experiment, SNR values are calculated for different methods. The results are presented in Table 3, and better ones are in bold. SNR is evaluated in dB, given by Eq. (9). Generally, the noise is considered imperceptible for human ears when its SNR is equal to or greater than 20 dB [38]. As shown in Table 3, the values of SNR for UAP are higher than those of FGSM. It infers the distortions introduced by UAP are slighter. For example, when ChenNet is considered the target detection network, the mean SNR value of UAP is about 29.64 dB, while the mean SNR value of FGSM is only 15.66 dB. The improvement of UAP is contributed to the constraint of the perturbation's magnitude as defined in Eq. (2).

Table 3: SNR values and time efficiency for different steganography methods against target detection networks

Detection networks	Steganography	Cover enhancement	0.5 bps		1.0 bps	
			SNR (dB)	Time (s)	SNR (dB)	Time (s)
FGSM [27]	LSBM	ChenNet	15.68	2,573.94	15.44	2,643.20
		LinNet	17.30	2,318.64	17.49	2,524.37
		BSNet	14.37	2,348.52	14.05	2,412.63
	LSBR	ChenNet	16.23	2,648.61	15.28	2,561.83
		LinNet	**17.27**	2,243.82	16.37	2,467.31
		BSNet	13.97	2,282.61	14.16	2,840.19
UAP (Proposed)	LSBM	ChenNet	**28.83**	**496.72**	**29.22**	**424.67**
		LinNet	**18.27**	**692.66**	**19.13**	**631.53**
		BSNet	**15.72**	**796.49**	**24.68**	**864.20**
	LSBR	ChenNet	**30.74**	**462.18**	**29.77**	**439.72**
		LinNet	17.06	**704.97**	**19.30**	**668.49**
		BSNet	**33.95**	**729.61**	**29.94**	**817.92**

4.4.4 Performance Evaluation of Time Efficiency

It is also significant to evaluate the time efficiency of constructing adversarial audios, especially for practical applications. In this experiment, the time efficiency of generating adversarial perturbation is considered. Table 3 presents the time cost of generating the perturbation on the training dataset for UAP and FGSM [27], respectively. It can be observed that the average time cost of UAP generation is almost 644.10 s, while the average time cost of FGSM perturbation generation is 2,488.81 s. It illustrates that the speed of UAP generation is much faster, nearly four times faster than the FGSM method. In addition, the proposed algorithm only needs to generate one single universal adversarial perturbation to deal with different testing samples at the evaluation stage, which is an important advantage in practical applications.

4.5 Performance Evaluation of Sample Diversification Strategy

Performances of sample diversification strategy (SDS) are evaluated in this section, which contains security performance evaluation and SNR evaluation.

4.5.1 Performance Evaluation of Security

In this subsection, the transferability of two approaches and their combination for SDS are evaluated, including noise addition and MP3 compression. The results of SDS with Gaussian white noises are first evaluated, and different noise factors are considered. The LSBM with 1.0 bps is considered for audio steganography. The results of security performance (P_m) are presented in Table 4, where noise factors are 0.002, 0.004, and 0.006, respectively. Networks used to generate the UAP (N_a) are listed in the second column, and networks used to evaluate the security (N_s) are listed in the first row. The better results are in bold, and the second best is underlined. It can be observed that SDS has

a positive influence on security in most cases. More specifically, the improvements of P_m are various for different combinations of the UAP generator and detection network. It should be noted that the security performance cannot be improved by simply increasing the magnitude of the noise factor.

Table 4: P_m with different noise factors in noise addition of SDS against unknown detection networks

Noise factor	N_a	N_s		
		ChenNet	LinNet	BSNet
Clean	ChenNet	–	0.0765	0.1722
	LinNet	0.4380	–	0.6078
	BSNet	0.3789	0.2660	–
0.002	ChenNet	–	**0.0843**	0.1852
	LinNet	**0.6215**	–	**0.7301**
	BSNet	0.3812	0.2680	–
0.004	ChenNet	–	0.0823	0.1980
	LinNet	0.6056	–	0.6749
	BSNet	**0.3887**	0.2780	–
0.006	ChenNet	–	0.0790	**0.2132**
	LinNet	0.5710	–	0.6813
	BSNet	0.2969	**0.2793**	–

Then, the P_m of applying SDS with the MP3 compression approach is presented in Table 5, where compression ratios are set as 1.0, 2.0, and 3.0, respectively. In general, SDS with MP3 compression also has a positive impact on security performance. Besides, it can be found that the values of P_m are higher when the compression ratio is set as 1.0 instead of 2.0 and 3.0 in most cases. It may be that the large compression ratio leads to more severe distortions of training samples which are harmful to the transferability. Therefore, the compression ratio is set to 1.0 for MP3 compression in this method. For the results in Tables 4 and 5, SDS with noise addition and MP3 compression can both be used to improve the transferability compared with the results in Table 2.

This paper also explores the results of SDS, considering the combination of noise addition and MP3 compression. In Table 6, the results of different versions of SDS are presented, including 1) SDS is not used, referred to as "Clean" (no noise addition and no MP3 compression); 2) Noise addition (with noise factor 0.002); 3) MP3 compression (with compression ratio 1.0); 4) Noise addition + MP3 compression (first conduct noise addition, then conduct MP3 compression); 5) MP3 compression + Noise addition (first conduct MP3 compression, then conduct noise addition) for attacking unknown detection networks. The best results are in bold, while the second-best results are underlined.

It can be observed that the values of P_m for different SDS methods are mostly better than those without sample diversification. Besides, the improvements of different SDS approaches are varied and related to the order of applying different processing operations, namely noise addition and MP3 compression. Specifically, most of the P_m values with "Noise addition + MP3 compression" are better than those of "MP3 compression + Noise addition". In fact, how to design the preprocessing operations to conduct the sample diversification is still an open issue in the field of adversarial attacks.

Table 5: P_m with different compression ratios in MP3 compression of SDS against unknown detection networks

Compression ratios	N_a	N_s		
		ChenNet	LinNet	BSNet
Clean	ChenNet	–	0.0765	0.1722
	LinNet	0.4380	–	0.6078
	BSNet	0.3789	0.2660	–
1.0	ChenNet	–	0.2973	**0.1766**
	LinNet	0.4059	–	0.7136
	BSNet	**0.3798**	**0.3373**	–
2.0	ChenNet	–	0.2780	0.1651
	LinNet	**0.4472**	–	**0.7368**
	BSNet	0.3794	0.2943	–
3.0	ChenNet	–	**0.3100**	0.1746
	LinNet	0.4433	–	0.7149
	BSNet	0.3312	0.3030	–

Table 6: P_m for attacking unknown detection networks with different sample diversification strategies (SDS)

Sample diversification strategy	N_a	N_s		
		ChenNet	LinNet	BSNet
Clean	ChenNet	–	0.0765	0.1722
	LinNet	0.4380	–	0.6078
	BSNet	0.3789	0.2660	–
Noise addition	ChenNet	–	0.0843	**0.1852**
	LinNet	**0.6215**	–	0.7301
	BSNet	0.3812	0.2680	–
MP3 compression	ChenNet	–	**0.2973**	0.1766
	LinNet	0.4059	–	0.7136
	BSNet	0.3785	**0.3373**	–
Noise addition + MP3 compression	ChenNet	–	0.2610	0.1717
	LinNet	0.4927	–	0.7295
	BSNet	**0.4288**	0.2487	–
MP3 compression + Noise addition	ChenNet	–	0.1823	0.1635
	LinNet	0.4845	–	**0.7463**
	BSNet	0.3255	0.2300	–

4.5.2 Performance Evaluation of SNR

The values of SNR for different SDS approaches are evaluated in Table 7. In most cases, SNR values are above or near 20 dB for different combinations of SDS processes. These results infer that the magnitude of the proposed UAPs is small and is not harmful to the quality of cover samples. It should be noted that UAPs generated by LinNet have lower SNR than those generated by ChenNet and BSNet since more iterations of LinNet are required for good transferability in black-box scenarios. Besides, applying SDS can improve the optimization efficiency of the adversarial perturbations, which leads to higher values of SNR compared with the adversarial perturbations generated by clean samples in most cases. It can be regarded as another advantage of the proposed SDS.

Table 7: SNR values for attacking unknown detection networks in sample diversification strategy

N_a	SDS				
	Clean	Noise addition	MP3 compression	Noise addition + MP3 compression	MP3 compression + Noise addition
ChenNet	29.21	29.59	39.08	38.13	33.56
LinNet	19.13	17.44	18.05	18.14	18.34
BSNet	24.67	25.17	24.85	25.39	25.43

4.6 Performance Evaluation of Perturbation Ensemble Method

Except for improving the transferability of adversarial perturbations by UAP and SDS, this paper also proposes an ensemble method, as shown in Section 4.6. In this experiment, the performance gain achieved by PEM is evaluated, and the results are presented in Table 8. Three different UAPs generated based on ChenNet, LinNet, and BSNet are used to obtain four types of ensembled UAPs. Besides, the SDS approach with "Noise addition + MP3 compression" is considered. The LSBM with 1.0 bps is used in this experiment. The results are presented in Table 8. The values of P_m for ensembled UAPs are higher than those of single UAPs in all cases, even when the ensembled UAP is used to attack an unknown network. For example, the results (clean SDS) of LinNet and BSNet to ChenNet are 0.4380 and 0.3789, as shown in Table 6, respectively, while the result of LinNet+BSNet to ChenNet is 0.5579, as shown in Table 8. It infers that the proposed PEM can fuse the features of UAPs generated by different networks to overcome the shortage of transferability for using a single UAP. It can also be observed that the performance improvement is various for different combinations of target networks. Due to the contribution of HPF in LinNet, perturbation combinations containing LinNet can achieve better performance improvements.

4.7 Comparison with other Steganography Methods

In this experiment, this paper compares the proposed method with other state-of-the-art steganography methods, including IA-SPP [13] and PixInWav [39]. More specifically, IA-SPP was designed to generate enhanced audios, which decomposed the perturbation at the point level and updated point-wise perturbations iteratively. PixInWav proposed a novel residual architecture operating on top of short-time discrete cosine transform audio spectrograms. In this experiment, the security performances (P_m) are evaluated by comparing different steganography methods against target detection networks, and the results are presented in Table 9. The LSBM with 1.0 bps is considered for audio steganography.

Table 8: P_m values for attacking unknown detection networks in the perturbation ensemble method (PEM)

PEM (N_a)	N_s			
	Sample diversification strategy	ChenNet	LinNet	BSNet
ChenNet + LinNet	Clean	0.5650	0.7757	0.6596
	Noise addition + MP3 compression	**0.5839**	**0.7780**	**0.7186**
LinNet + BSNet	Clean	0.5579	0.7920	0.6686
	Noise addition + MP3 compression	**0.5937**	**0.7967**	**0.7465**
ChenNet + BSNet	Clean	0.3473	0.5302	0.5519
	Noise addition + MP3 compression	**0.4397**	**0.5593**	**0.6526**

Table 9: Security performances (P_m) by comparing different steganography methods

Steganography	ChenNet	LinNet	BSNet
IA-SPP [13]	0.8512	0.7785	0.6997
PixInWav [39]	0.7922	0.7476	0.6410
UAP (Proposed)	**0.8906**	**0.8220**	**0.7618**

As shown in Table 9, the proposed method can still achieve outstanding security performance for different detection networks. Compared with other steganography methods like IA-SPP and PixInWav, the proposed method can obtain the largest value of P_m, which infers better security of the proposed UAP. Besides, it can be observed the improvement of the proposed method for BSNet is more discriminative, which may be due to the fact that the detection capability of BSNet is stronger and leads to a more powerful UAP than ChenNet and LinNet.

This paper also evaluates the transferability of different methods against different detection networks. In this comparison experiment, only IA-SPP is considered since IA-SPP applied adversarial perturbations to enhance the secret audios, which is similar to this work. However, PixInWav focuses on constructing a network to generate secret audios, which cannot conduct black-box attack scenarios. For a fair comparison, PEM is applied to all methods. As shown in Table 10, the proposed method can still achieve outstanding transferability performance. Compared with the "Baseline" method, the proposed method and IA-SPP can achieve distinct performance gains by applying adversarial perturbations. Since IA-SPP belongs to a post-processing method of secret audios, whose perturbations may destroy steganographic information, its results are much worse than those of the proposed method. In summary, the proposed method can still achieve promising performance compared with other state-of-the-art methods.

Table 10: Transferability performances (P_m) by comparing different steganography methods

Steganography	ChenNet	LinNet	BSNet
Baseline (no enhancement)	0.0148	0.1239	0.0472
IA-SPP [13]	0.4862	0.5764	0.6410
Proposed	**0.6525**	**0.7993**	**0.7494**

4.8 The Influence of Different Attack Success Rates

In this experiment, the influence of different desired success rates is evaluated. In the proposed method, the desired success rate is the threshold of Attack Success Rate (ASR). Specifically, diversified secret audios in the training subset were obtained by clean SDS, and LSBM steganography is considered with 1.0 bps. The desired success rate is set from 0.1 to 0.9 with step 0.1, and other parameters are set as default. Fig. 5 presents the results of ASR and SNR with different rates and detection networks. It is observed that the values of ASR become higher with the increment of the desired success rate, while the larger rate can cause lower SNR (worse quality of audio samples). Different detection networks have similar tendencies, except that the ASR of ChenNet can achieve a distinct improvement when the rate is 0.4 because of the increment of iterations. Therefore, the desired success rate is set to 0.9 in this method, considering both the distortion introduced by the perturbation and attack performance.

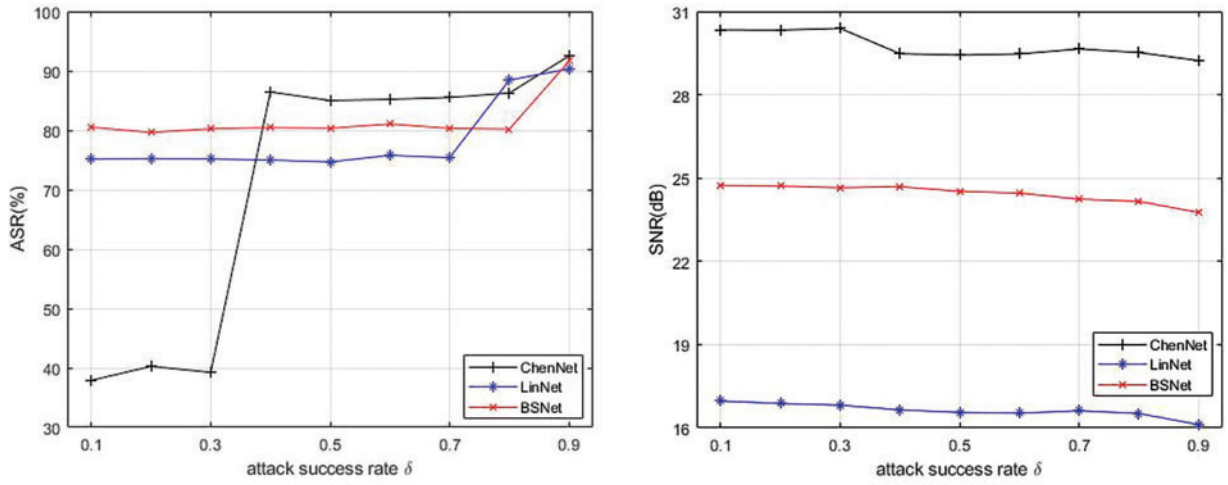

Figure 5: The effect of different desired success rates on the values of ASR (a) and SNR (b) against different detection networks

5 Conclusion

In this work, this paper proposed a cover enhancement method for audio steganography based on universal adversarial perturbations with sample diversification. The universal adversarial perturbation is iteratively constructed based on the adversarial example technique of Deepfool, aiming for the specific attack success rate as the optimization objective. In addition, the sample diversification

strategy and perturbation ensemble method are designed to improve the transferability of adversarial perturbations in black-box attack scenarios.

Extensive experimental results demonstrate that both noise addition and MP3 compression, two audio processing operations in SDS, contribute to the improvements of transferability against unknown detection networks. Besides, the security performances of applying ensembled UAPs via PEM are better than those with a single UAP in black-box attacks. When the cover enhancement method combines SDS and PEM, the proposed scheme can achieve much better security performances of audio steganography methods than their original versions. It is very convenient to equip the proposed audio cover enhancement method with existing audio steganography methods, which is an advantage in practical applications. In future work, the adaptive weighting methods for perturbation ensembles will be explored, and more advanced strategies to improve the transferability of adversarial perturbations will be considered.

Funding Statement: This work was supported by the National Natural Science Foundation of China (61902263) and the National Key Research and Development Program of China (2018YFB0804103).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. M. Ghadirli, A. Nodehi and R. Enayatifar, "An overview of encryption algorithms in color images," *Signal Processing*, vol. 164, no. 2018, pp. 163–185, 2019.

[2] B. Li, Y. Feng, Z. Xiong, W. Yang and G. Liu, "Research on AI security enhanced encryption algorithm of autonomous IoT systems," *Information Sciences*, vol. 575, no. 4, pp. 379–398, 2021.

[3] A. Anand and A. K. Singh, "Watermarking techniques for medical data authentication: A survey," *Multimedia Tools and Applications*, vol. 80, no. 20, pp. 30165–30197, 2021.

[4] J. Khandelwal, V. Kumar Sharma, D. Singh and A. Zaguia, "DWT-SVD based image steganography using threshold value encryption method," *Computers, Materials & Continua*, vol. 72, no. 2, pp. 3299–3312, 2022.

[5] K. Manjunath, G. K. Ramaiah and M. GiriPrasad, "Backward movement oriented shark smell optimization-based audio steganography using encryption and compression strategies," *Digital Signal Processing*, vol. 122, no. 25, pp. 103335, 2022.

[6] R. Li, J. Qin, Y. Tan and N. N. Xiong, "Coverless video steganography based on frame sequence perceptual distance mapping," *Computers, Materials & Continua*, vol. 73, no. 1, pp. 1571–1583, 2022.

[7] R. Sridevi, A. Damodaram and S. Narasimham, "Efficient method of audio steganography by modified LSB algorithm and strong encryption key with enhanced security," *Journal of Theoretical & Applied Information Technology*, vol. 5, no. 6, pp. 768–771, 2009.

[8] S. Rekik, D. Guerchi, S. Selouani and H. Hamam, "Speech steganography using wavelet and Fourier transforms," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2012, no. 1, pp. 1–14, 2012.

[9] B. Chen, W. Luo and H. Li, "Audio steganalysis with convolutional neural network," in *Proc. of the 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, PA, USA, pp. 85–90, 2017.

[10] Y. Lin, R. Wang, D. Yan, L. Dong and X. Zhang, "Audio steganalysis with improved convolutional neural network," in *Proc. of the ACM Workshop on Information Hiding and Multimedia Security*, Paris, France, pp. 210–215, 2019.

[11] D. Lee, T. Oh and K. Kim, "Deep audio steganalysis in time domain," in *Proc. of the 2020 ACM Workshop on Information Hiding and Multimedia Security*, Denver, CO, USA, pp. 11–21, 2020.

[12] J. Wu, B. Chen, W. Luo and Y. Fang, "Audio steganography based on iterative adversarial attacks against convolutional neural networks," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2282–2294, 2020.

[13] K. Ying, R. Wang and D. Yan, "Iteratively generated adversarial perturbation for audio stego post-processing," in *2021 IEEE Int. Workshop on Information Forensics and Security (WIFS)*, Montpellier, France, pp. 1–6, 2021.
[14] S. Moosavi-Dezfooli, A. Fawzi and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 2574–2582, 2016.
[15] S. Roy, J. Parida, A. K. Singh and A. S. Sairam, "Audio steganography using LSB encoding technique with increased capacity and bit error rate optimization," in *Proc. of the Second Int. Conf. on Computational Science*, Coimbatore, India, Engineering and Information Technology, pp. 372–376, 2012.
[16] M. A. Ahmed, M. M. Kiah, B. Zaidan and A. Zaidan, "A novel embedding method to increase capacity and robustness of low-bit encoding audio steganography technique using noise gate software logic algorithm," *Journal of Applied Sciences*, vol. 10, no. 1, pp. 59–64, 2010.
[17] H. Kumar and A. Taluja, "Enhanced LSB technique for audio steganography," in *2012 Third Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT'12)*, Coimbatore, India, pp. 1–4, 2012.
[18] D. C. Kar and C. J. Mulkey, "A multi-threshold based audio steganography scheme," *Journal of Information Security and Applications*, vol. 23, no. 3–4, pp. 54–67, 2015.
[19] W. Luo, Y. Zhang and H. Li, "Adaptive audio steganography based on advanced audio coding and syndrome-trellis coding," in *Int. Workshop on Digital Watermarking*, Magdeburg, Germany, pp. 177–186, 2017.
[20] K. Chen, H. Zhou, W. Li, K. Yang, W. Zhang *et al.,* "Derivative-based steganographic distortion and its non-additive extensions for audio," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2027–2032, 2019.
[21] Q. Liu, A. H. Sung and M. Qiao, "Derivative-based audio steganalysis," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 7, no. 3, pp. 1–19, 2011.
[22] W. Luo, H. Li, Q. Yan, R. Yang and J. Huang, "Improved audio steganalytic feature and its applications in audio forensics," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 2, pp. 1–14, 2018.
[23] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.
[24] Z. Zhang, X. Yi and X. Zhao, "Improving audio steganalysis using deep residual networks," in *Int. Workshop on Digital Watermarking*, Chengdu, China, pp. 57–70, 2019.
[25] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan *et al.,* "Intriguing properties of neural networks," arXiv preprint arXiv: 1312.6199, 2013.
[26] Y. Zhang, W. Zhang, K. Chen, J. Liu, Y. Liu *et al.,* "Adversarial examples against deep neural network based steganalysis," in *Proc. of the 6th ACM Workshop on Information Hiding and Multimedia Security*, Innsbruck, Austria, pp. 67–72, 2018.
[27] I. J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd Int. Conf. on Learning Representations*, San Diego, CA, USA, pp. 1–11, 2015.
[28] L. Zhou, G. Feng, L. Shen and X. Zhang, "On security enhancement of steganography via generative adversarial image," *IEEE Signal Processing Letters*, vol. 27, pp. 166–170, 2019.
[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley *et al.,* "Generative adversarial nets," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[30] C. Qin, W. Zhang, X. Dong, H. Zha and N. Yu, "Adversarial steganography based on sparse cover enhancement," *Journal of Visual Communication and Image Representation*, vol. 80, no. 2, pp. 103325, 2021.
[31] L. Chen, R. Wang, D. Yan and J. Wang, "Learning to generate steganographic cover for audio steganography using GAN," *IEEE Access*, vol. 9, pp. 88098–88107, 2021.
[32] J. Mielikainen, "LSB matching revisited," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 285–287, 2006.

[33] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang *et al.,* "Improving transferability of adversarial examples with input diversity," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 2730–2739, 2019.
[34] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi and P. Frossard, "Universal adversarial perturbations," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 1765–1773, 2017.
[35] S. Abdoli, L. G. Hafemann, J. Rony, I. B. Ayed, P. Cardinal *et al.,* "Universal adversarial audio perturbations," arXiv preprint arXiv: 1908.03173, 2019.
[36] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1. 1," *NASA STI/Recon Technical Report*, vol. 93, pp. 27403, 1993.
[37] C. Chan and L. Cheng, "Hiding data in images by simple LSB substitution," *Pattern Recognition*, vol. 37, no. 3, pp. 469–474, 2004.
[38] T. Du, S. Ji, J. Li, Q. Gu, T. Wang *et al.,* "Sirenattack: Generating adversarial audio for end-to-end acoustic systems," in *Proc. of the 15th ACM Asia Conf. on Computer and Communications Security*, Taipei, Taiwan, pp. 357–369, 2020.
[39] M. Geleta, C. Punti and K. McGuinness, "PixInWav: Residual steganography for hiding pixels in audio," in *ICASSP 2022-2022 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, pp. 2485–2489, 2022.